St. Jude Children's Research Hospital

“BIG Pediatric Cancer Genomic Data: Discovery, Precision Medicine, and Data Sharing”

Jinghui Zhang

Member and Chair

Department of Computational Biology

St. Jude Endowed Chair in Bioinformatics

St. Jude Graduate School of Biomedical Sciences

St. Jude Children's Research Hospital

St. Jude Children's Research Hospital

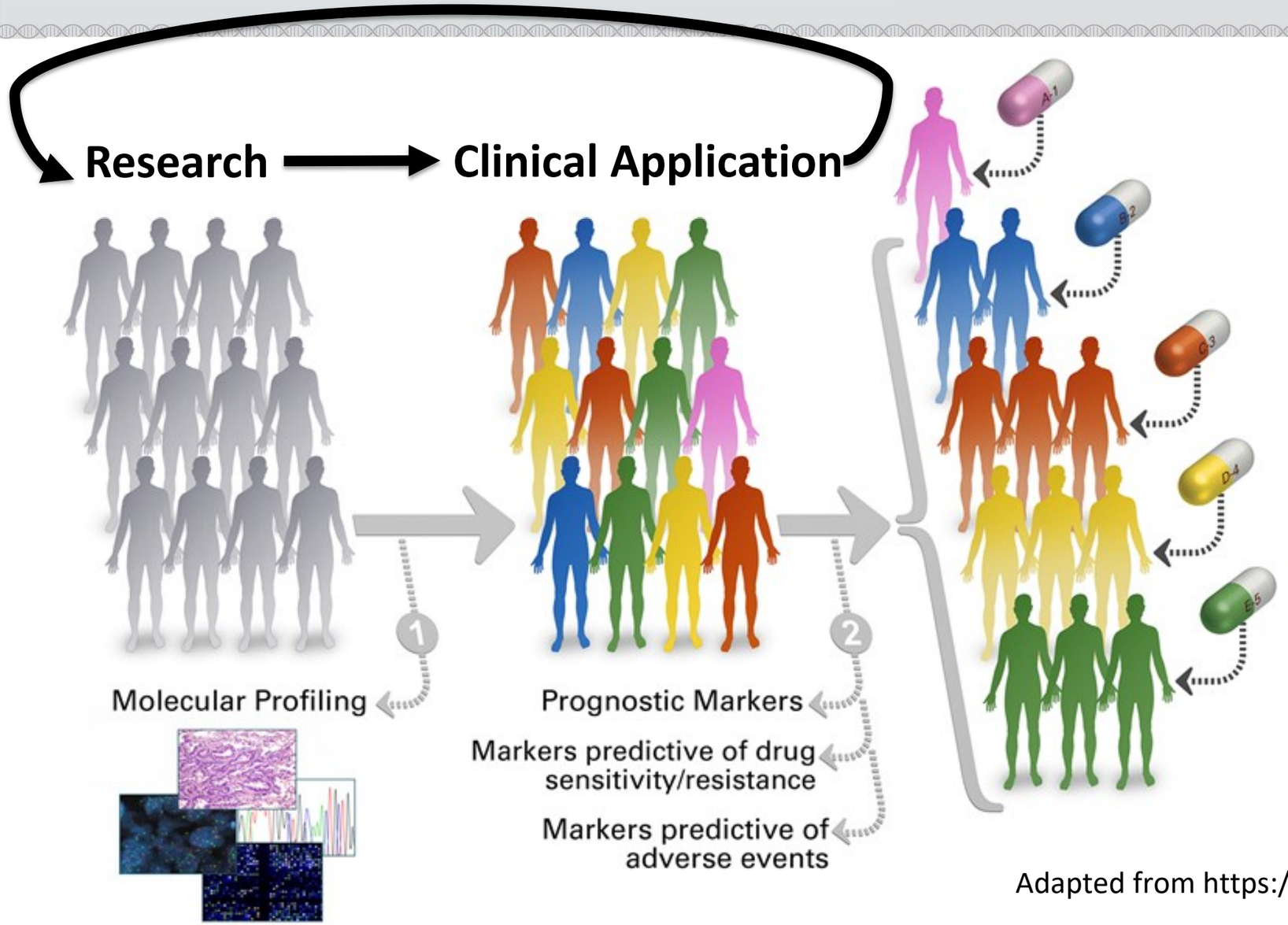
BIG Pediatric Cancer Genomic Data: Discovery, Precision Medicine, and Data Sharing

Jinghui Zhang, PhD

Chair, Member

Department of Computational Biology

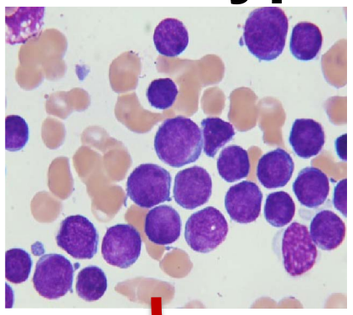
BIG Data & Precision Medicine



Adapted from https://pct.mdanderson.org/#/

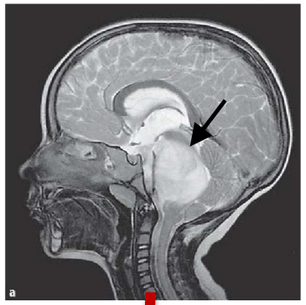
Pediatric Cancer Genome Project (PCGP) 2010-2013

Leukemia
12 Subtypes



Solid Tumors
7 Subtypes

Brain Tumors
5 Subtypes

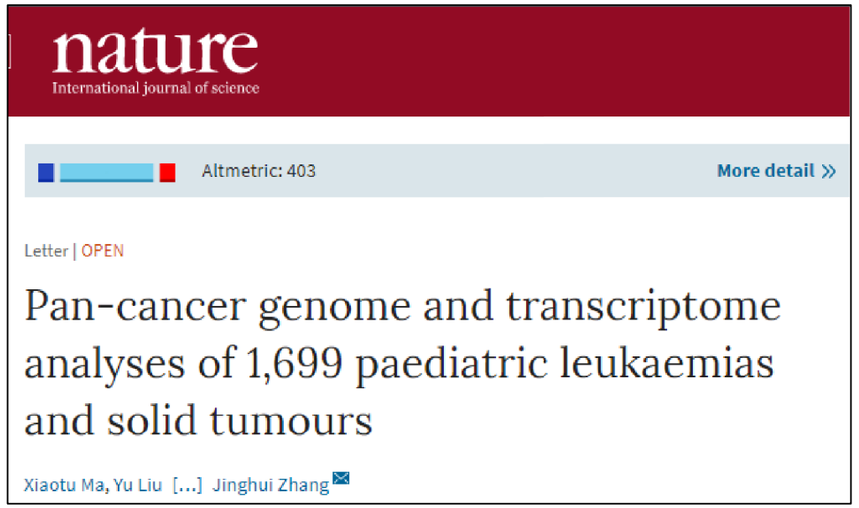


700 Tumor/Normal WGS Pairs
1500 WES & 1000 RNA-seq
>2,000,000 lesions verified

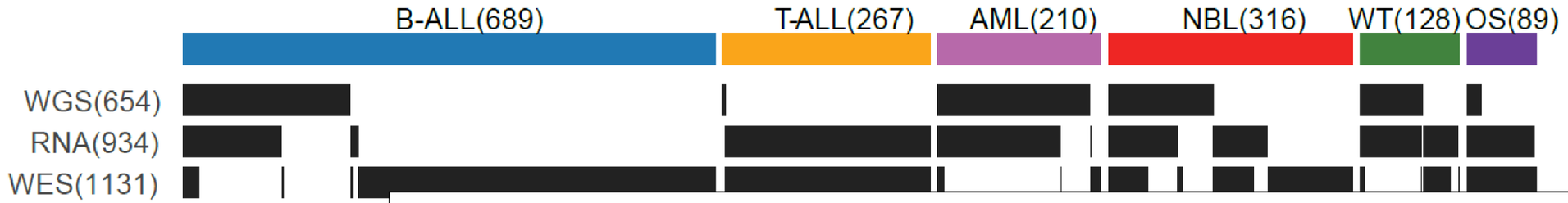
35 High-Impact Published Studies on Pediatric Cancer Driver Genes

Pan-cancer Study of NCI TARGET

Daniela S. Gerhard
Stephen P. Hunger (ALL)
Soheil Meshinchi (AML)
John M. Maris (Neuroblastoma)
Elizabeth J. Perlman (Wilms Tumor)
Ching C. Lau (Osterosarcoma)
Paul S. Meltzer (Osterosarcoma)
TARGET Analysis Working Group (TAWG)



Pan-cancer Analysis of WGS, WES and RNA-seq of 1,699 Patient Samples

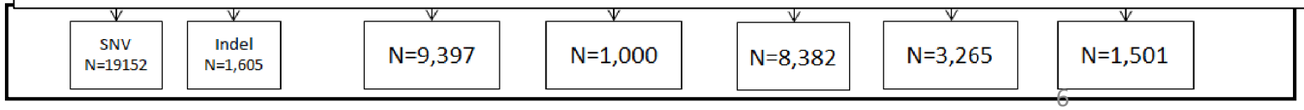


100 million candidate variants

coding variants=44,302

Of the 142 statistically significant driver genes:

- 78 (55%) absent in three adult pan-cancer studies
  - Kandoth et al Nature (2013)
  - Lawrence et al Nature (2013)
  - Zack et al Nature Genetics (2013)
- 62% are SVs, CNVs

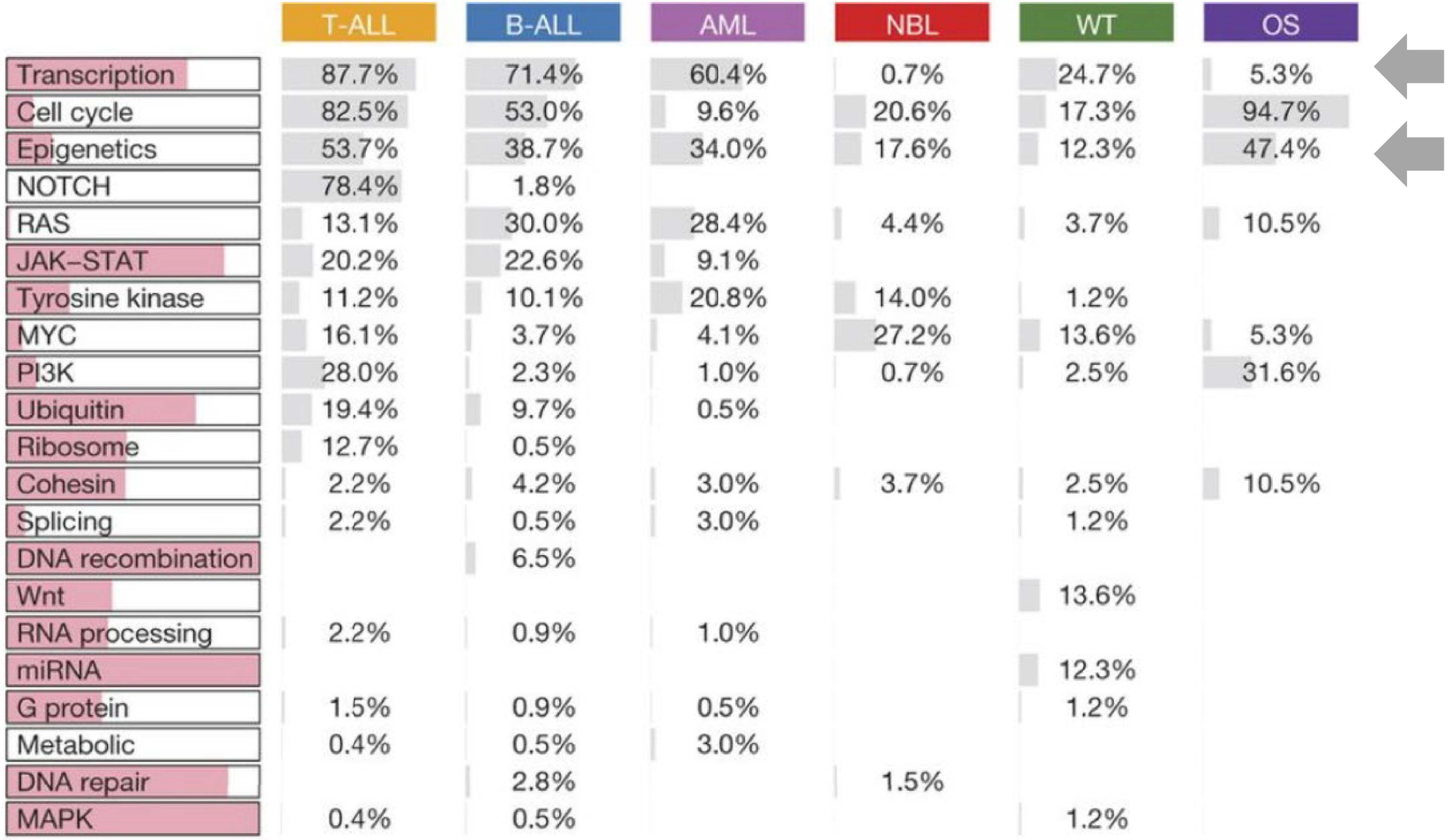


142 driver genes ← Statistical analysis (MutSigCV, GRIN) ← Pathogenicity classification → 82 additional driver genes with P/LP variants

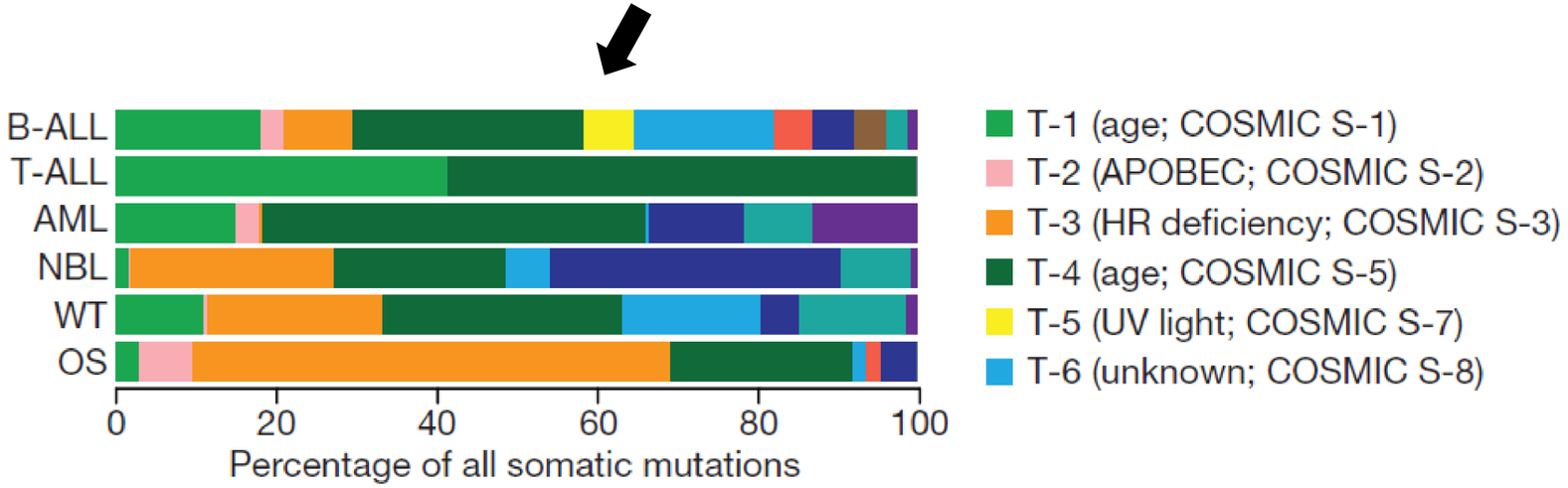
Biological Processes Altered in Pediatric Cancer

Proportion of Genes Unique to Pediatric Cancer

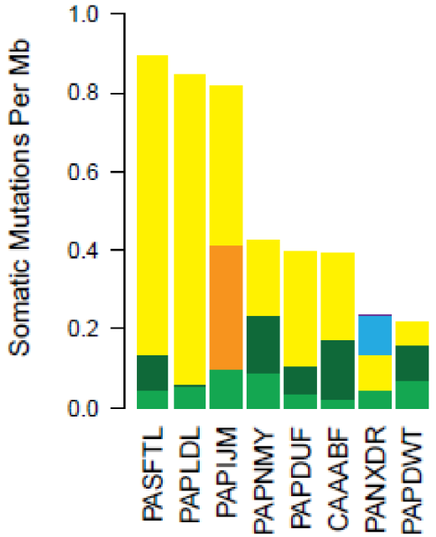
Proportion of Genes Shared with Adult Cancer



Mutational Signatures of Pediatric Cancers



Ludmil Alexandrov



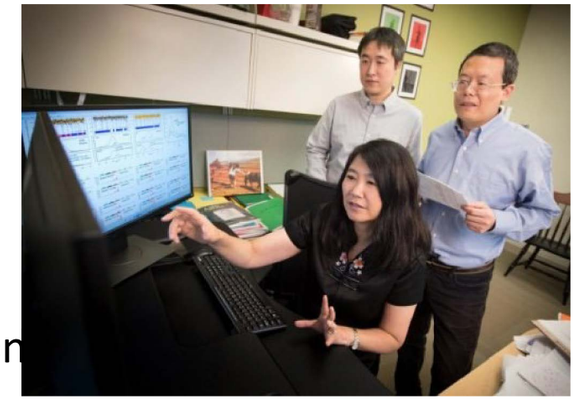
List of samples with UV signatures

USI	Gender	Race	Age at Dx	ALL subtype
PAPLDL	Female	Hispanic	7	iAMP21
PASFTL	Female	Caucasian	4	Hyperdiploid
PAPIJM	Male	Caucasian	11	Hyperdiploid
PAPDUF	Male	Caucasian	3	Hyperdiploid
PAPNMY	Male	Caucasian	14	iAMP21
CAAABF	Male	Caucasian*	3	Hypodiploid
PANXDR	Male	Caucasian*	3	Hyperdiploid
PAPDWT	Female	Asian	6	Hyperdiploid

*: inferred from genomics data

Acknowledgement

- **Zhang Lab**
  - Xiaotu Ma
  - Yu Liu
  - Yanling Liu
  - Xin Zhou
  - Yongjin Li
  - Michael Edmonson



- **NCI TARGET Team**
  - Daniela S. Gerhard
  - Steve Hunger
  - Soheil Meshinchi
  - John Maris
  - Ching C. Lau
  - Paul S. Meltzer
  - TARGET Analysis Working Group (TAWG)

- **Gawad Lab**
  - Chuck Gawad
  - Veronica Gonzalez-Pena
- **Comp. Bio. Genomics Lab**
  - John Easton
  - Li Dong
- **Comp. Bio. Software Group**
  - Michael Rusch
  - Mark Wilkinson
  - Edgar Sioson
- **Bio. Stat. Department**
  - Stan Pounds
  - Xueyan Cao

- **Other Collaborators**
  - Ludmil Alexandrov (UCSD)
  - Robert Huether (Tempus)

St. Jude Children's Research Hospital

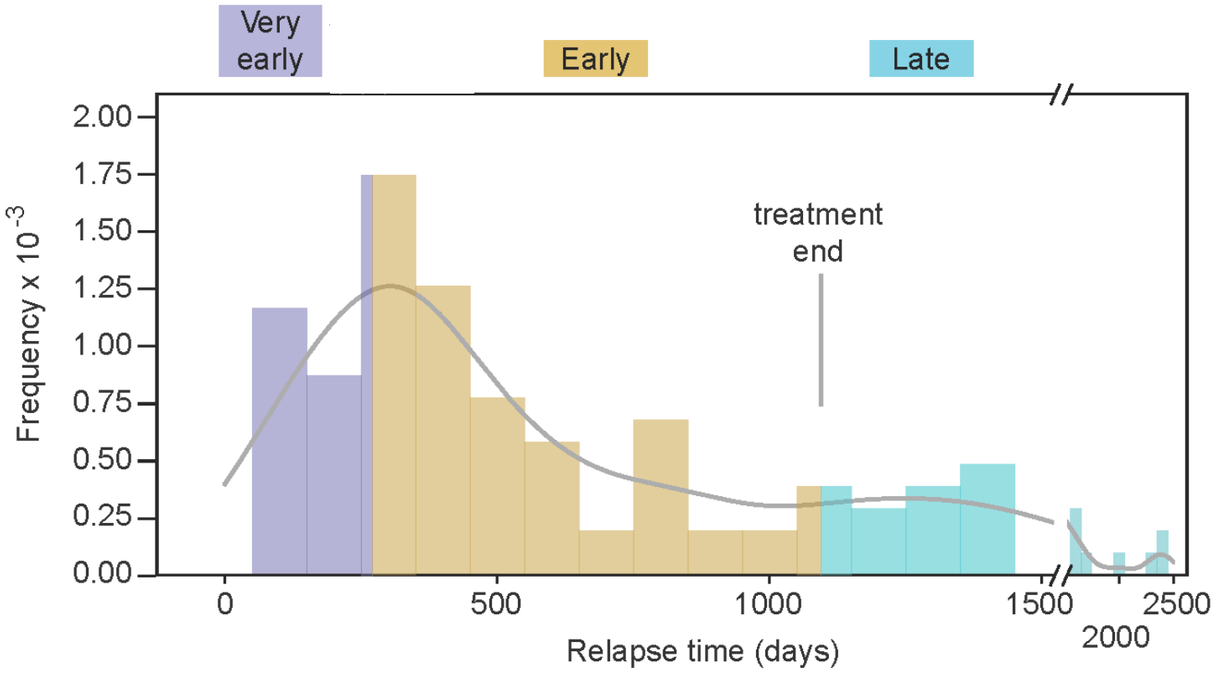
Drug Resistant Mutations in Relapsed Pediatric Acute Lymphoblastic Leukemia

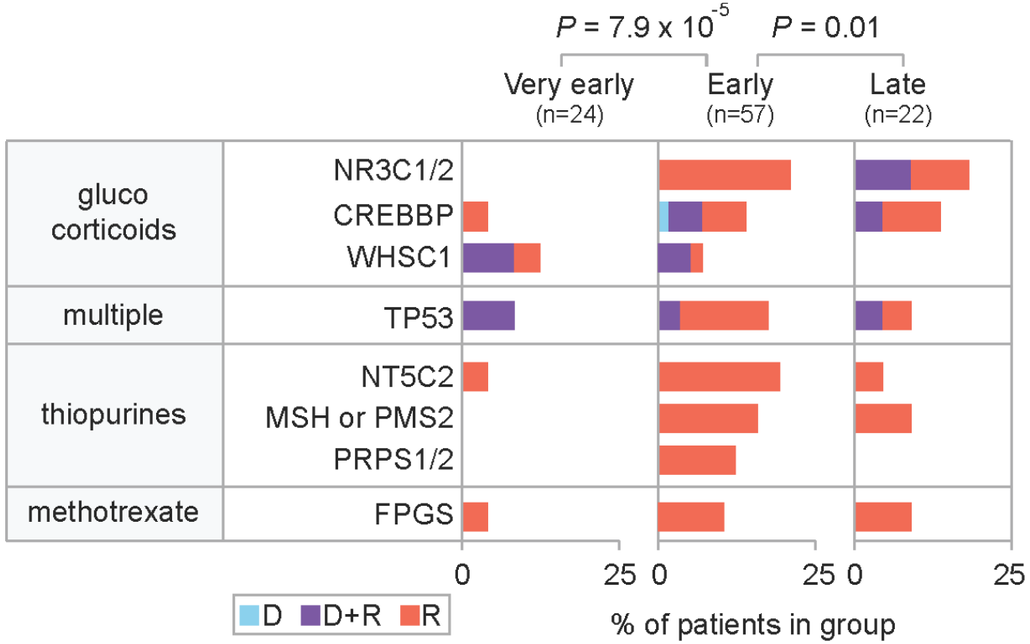
A collaboration with Shanghai Children's Medical Center (SCMC)

Relapsed Acute Lymphoblastic Leukemia (ALL)

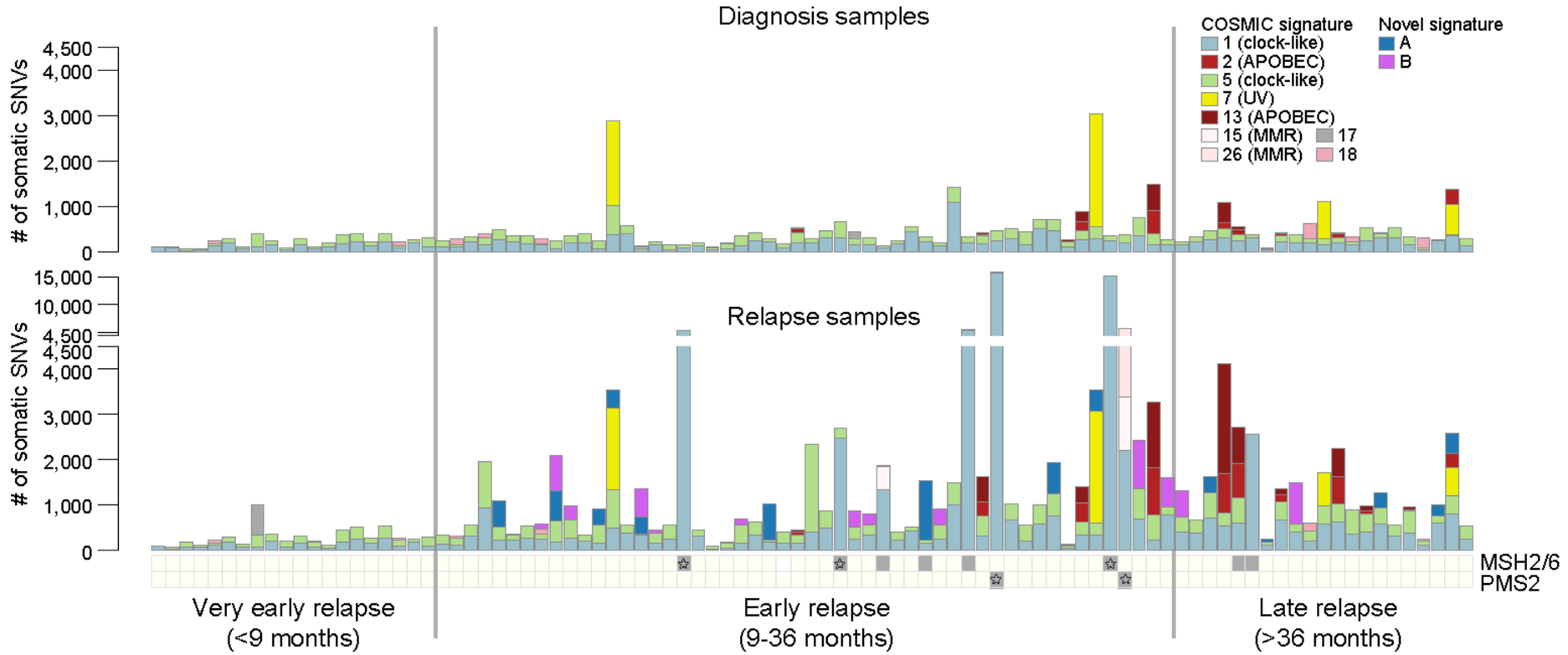
Li, Brady et al, Blood 2020

- 103 patients with very early, early and late relapse were analyzed by WGS and RNA-seq of diagnosis (D)-relapse (R)-germline trio
- Relapse-specific mutations enriched in 12-genes known to be involved in drug response

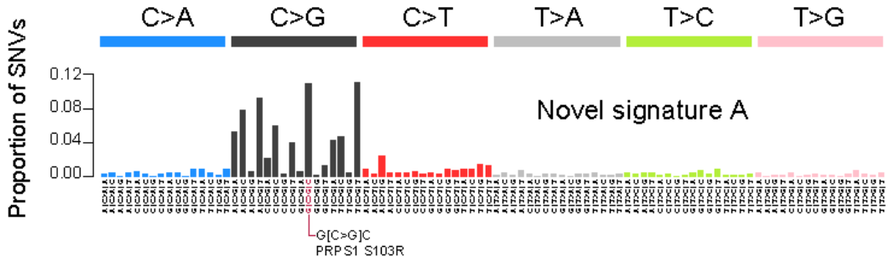




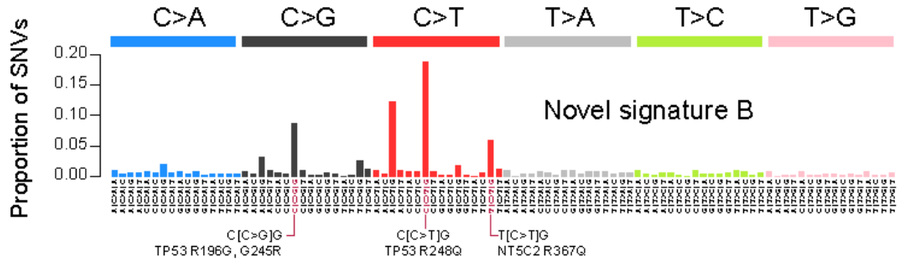
Two Relapse-specific Novel Signatures



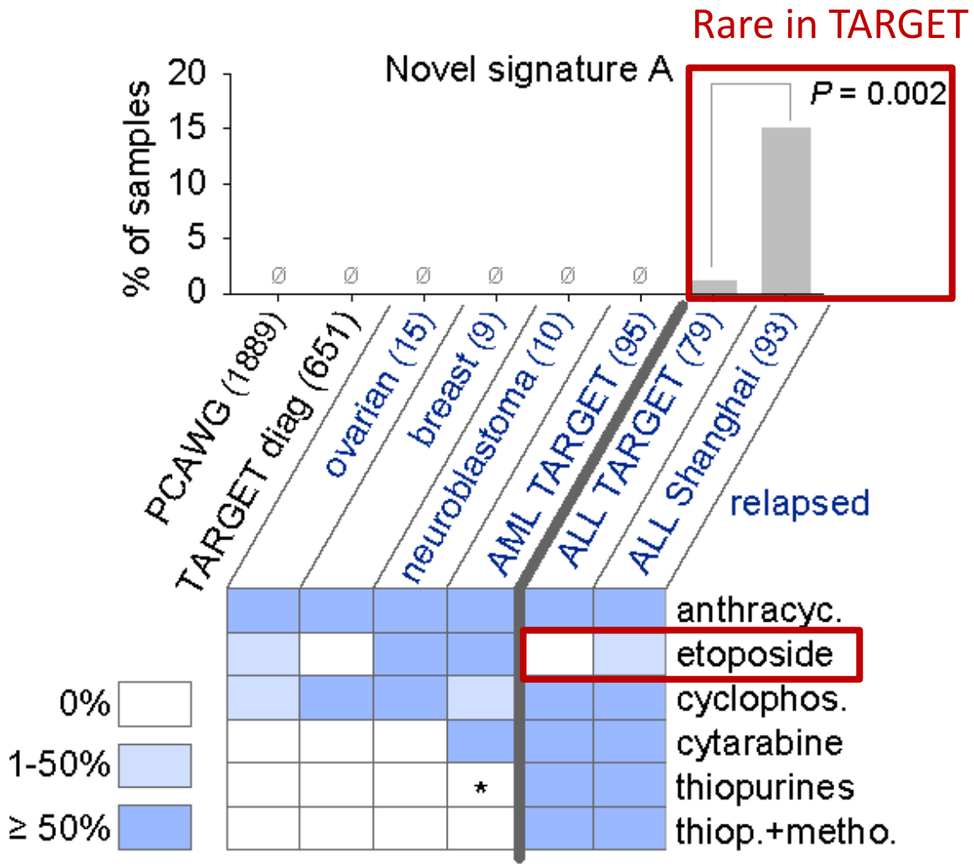
Novel A (PRPS1, TP53)

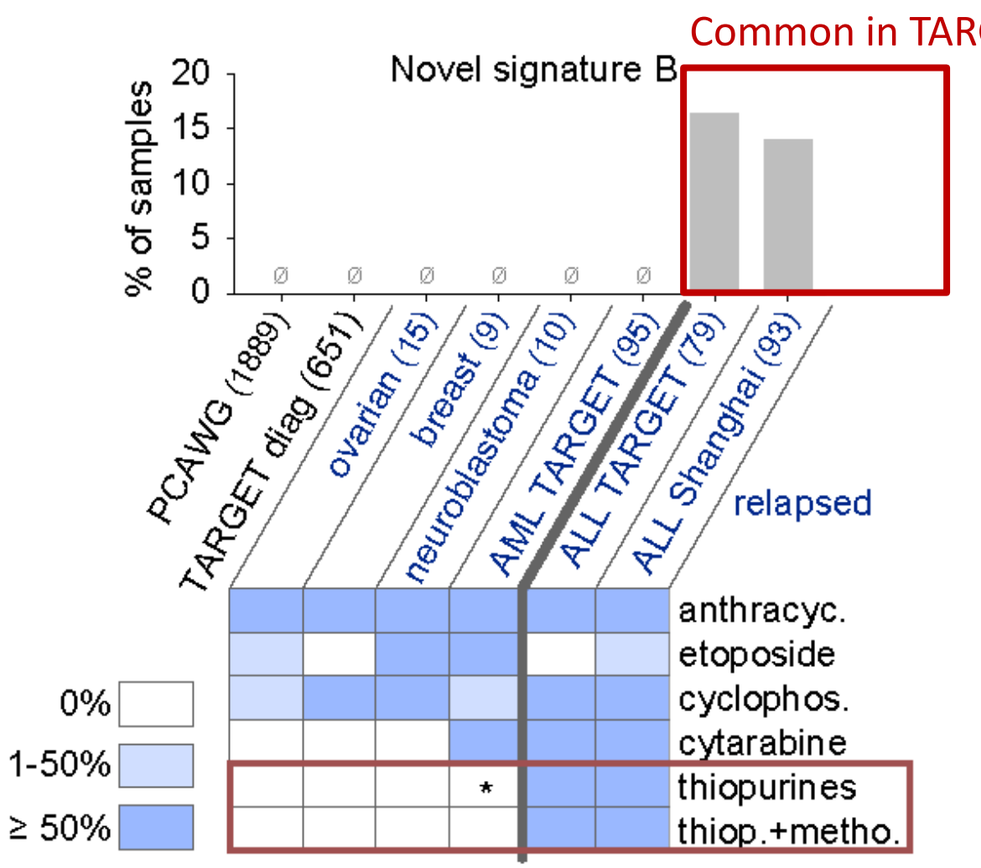


Novel B (NT5C2, NR3C1, TP53)



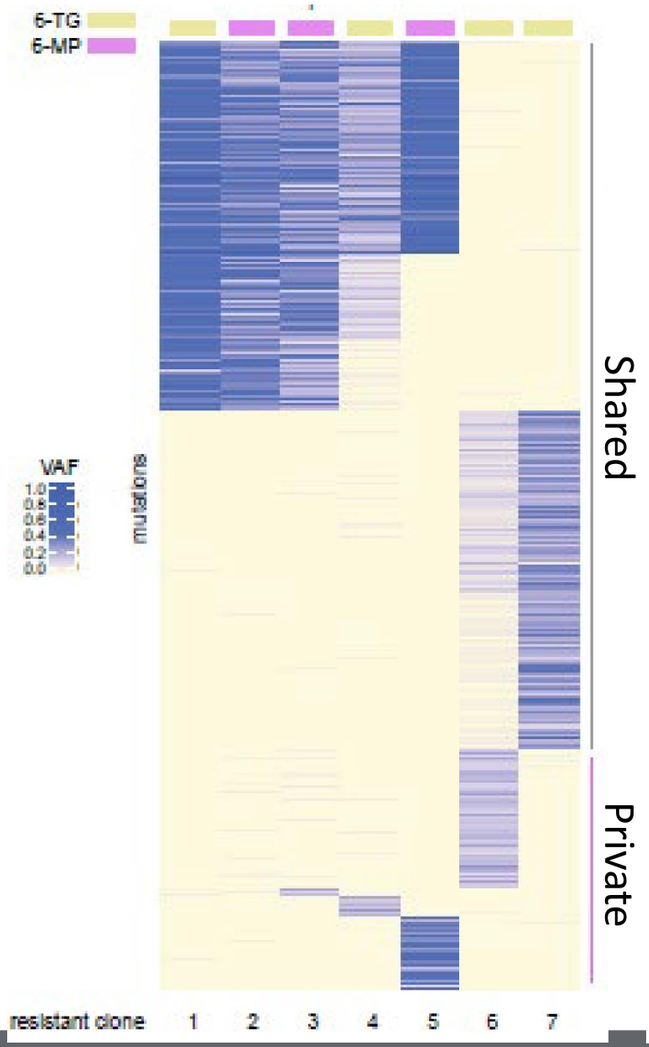
Candidate Mutagenic Agents for Novel Signatures



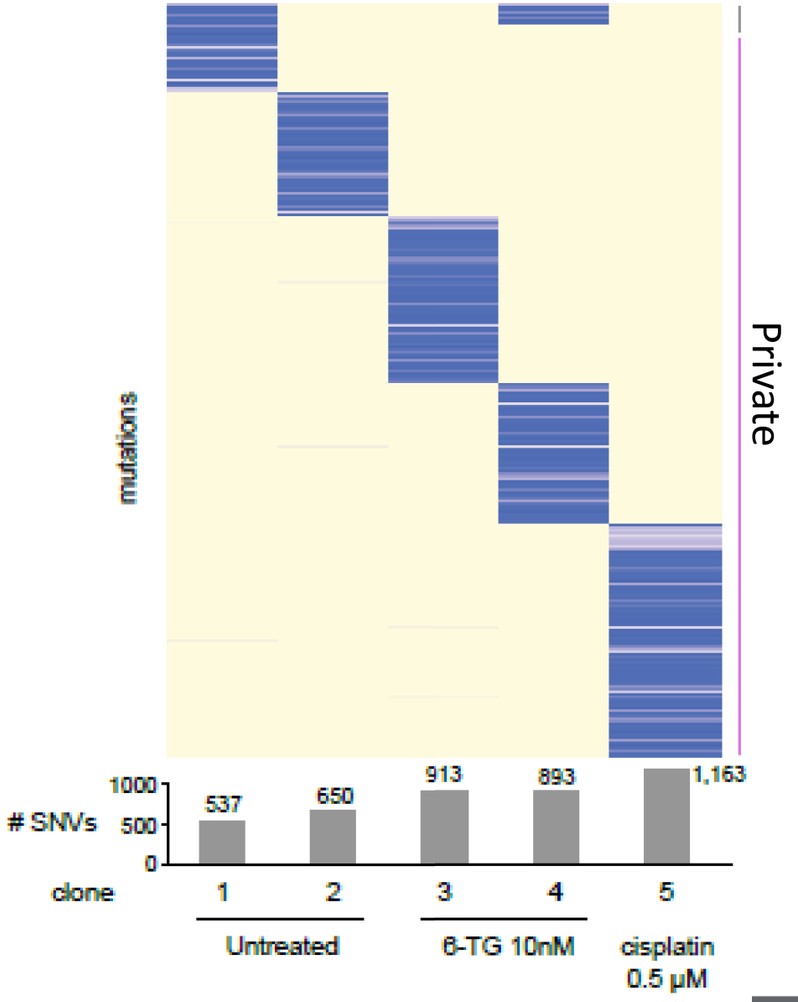


Experimental Confirmation of Thiopurine-induced Novel Signature B

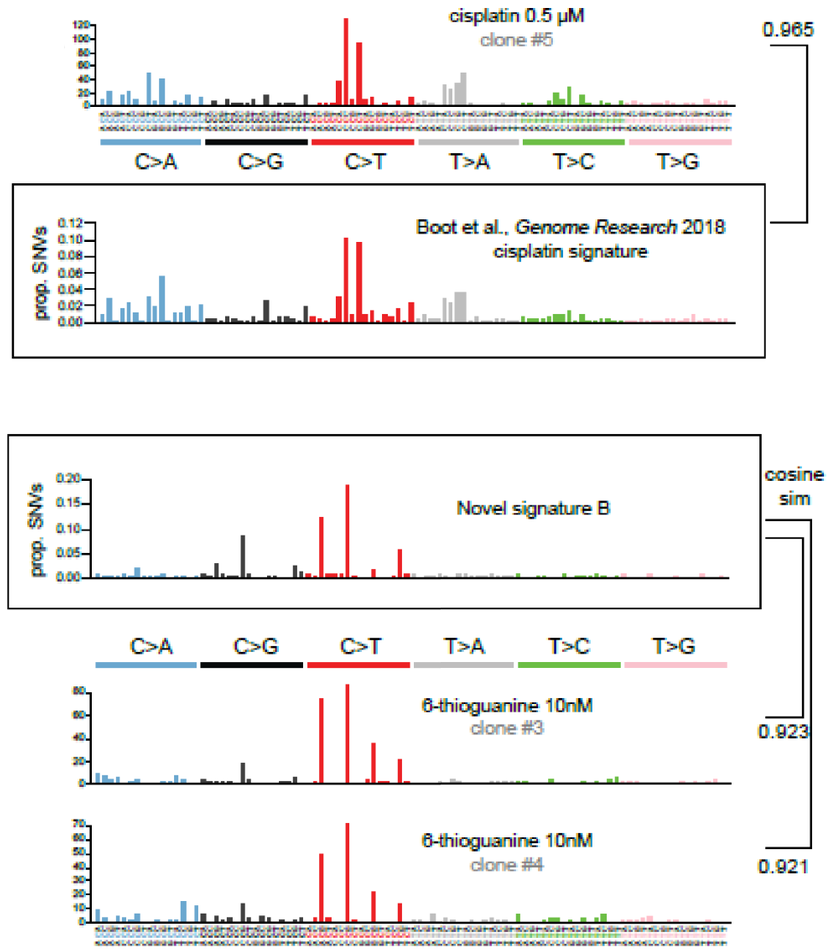
ALL Cell line REH



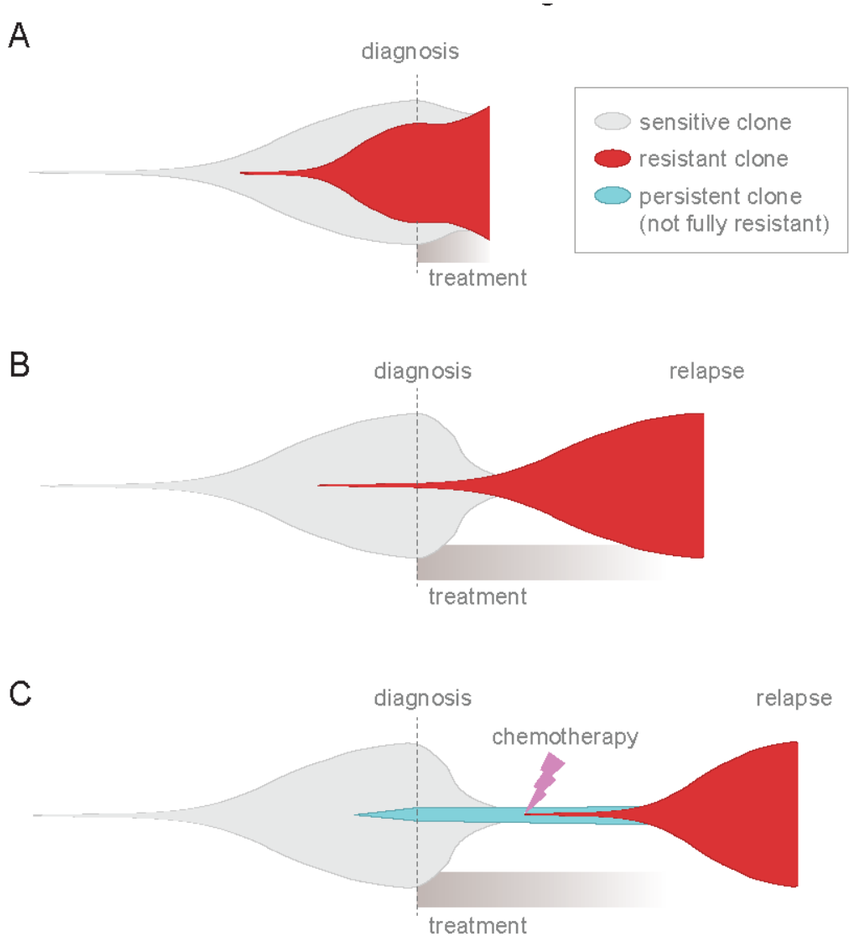
Normal Breast Cell line MCF10A



Signature of MCF10A Mutations



A New Model for ALL Relapse



***de novo* resistance**
not observed in ALL

chemo-selection
very early relapses

New Model
Persistent clones with chemo-induced mutations
early & late relapses

Li, Brady, Ma, et al, Blood, 2020

Acknowledgement

- Shanghai Children's Medical Center
  - Benshang Li
  - Shuhong Shen
  - Jingyan Tang
  - Bin-bing Zhou
- Tianjin Institute of Hematology
  - Yingchi Zhang
  - Xiaofan Zhu
- Anhui Medical University
  - Ningling Wang
- St Jude

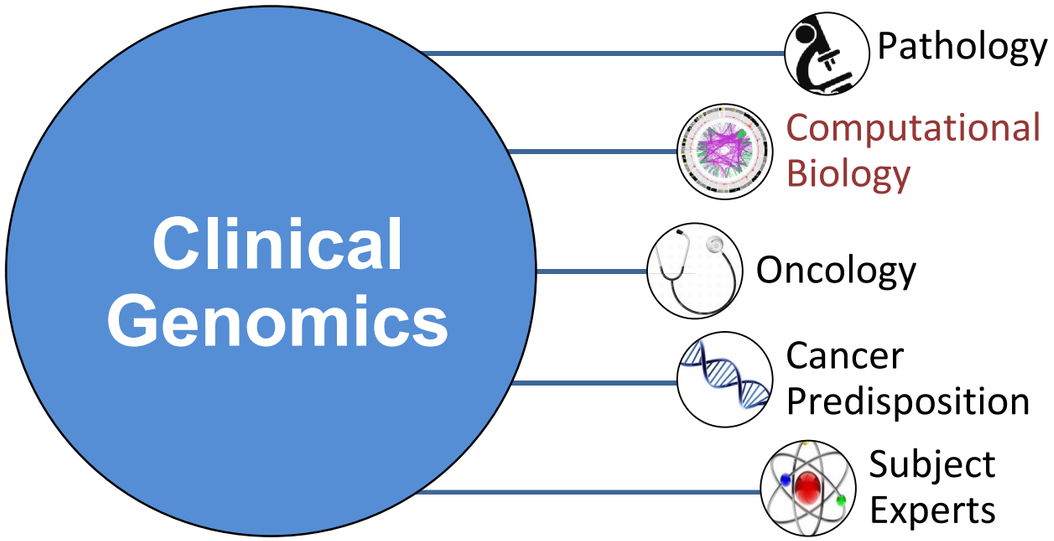
  - **Sam Brady**
  - Yongjin Li
  - Xiaotu Ma
  - Yu Liu
  - James Downing
  - Ching-Hon Pui
  - Jun J. Yang
  - Jinghui Zhang
  - 9 Additional Scientists
- Princeton University
  - Matthew Myers
  - Ben Raphael

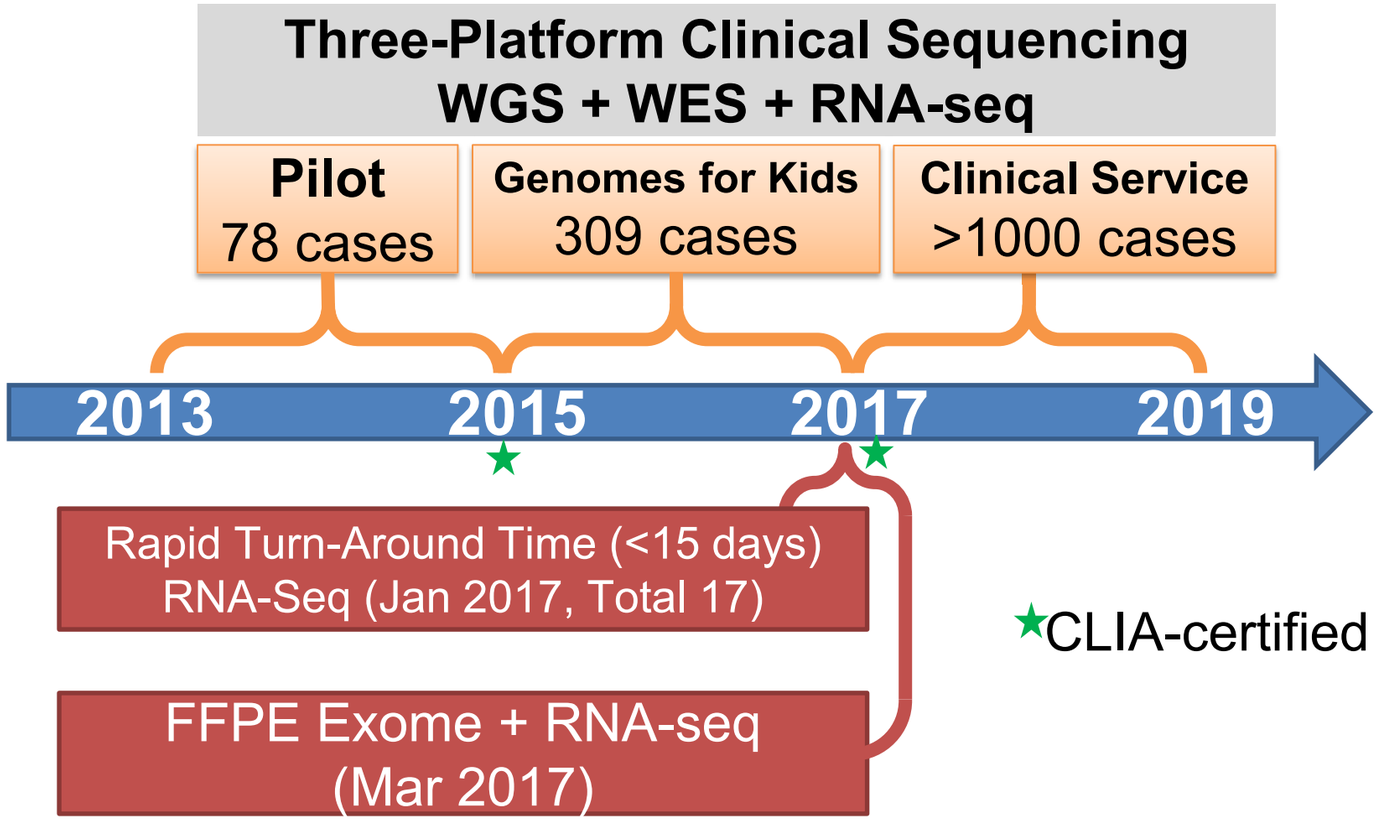
Li, Brady, Ma, et al, Blood, 2020

St. Jude Children's Research Hospital

Clinical Genomics

Clinical Genomics: Timeline





ClinGen Pipeline for 3-Platform Sequencing

Pipelines developed from 78 cases in a pilot study

Three-platform Sequencing → Variant Detection, Cross-Validation → Variant Classification → Analyst Curation, Panel Review → Report Generation

Tumor

30X
Whole Genome
•
100X
Whole Exome
•
RNA-Seq

Normal

30X
Whole Genome
•
100X
Whole Exome

Somatic SNV
Germline SNV
Somatic Indel
Germline Indel
Somatic CNA
Germline CNA
Somatic SV
Somatic Fusion
Chromothripsis
LOH
Ploidy
Splicing
Expression
Purity
Contamination
Mosaicism

PCGP
COSMIC
ClinVar
Other Databases
Literature

Pathogenic
Likely Pathogenic
VUS
Likely Benign
Benign

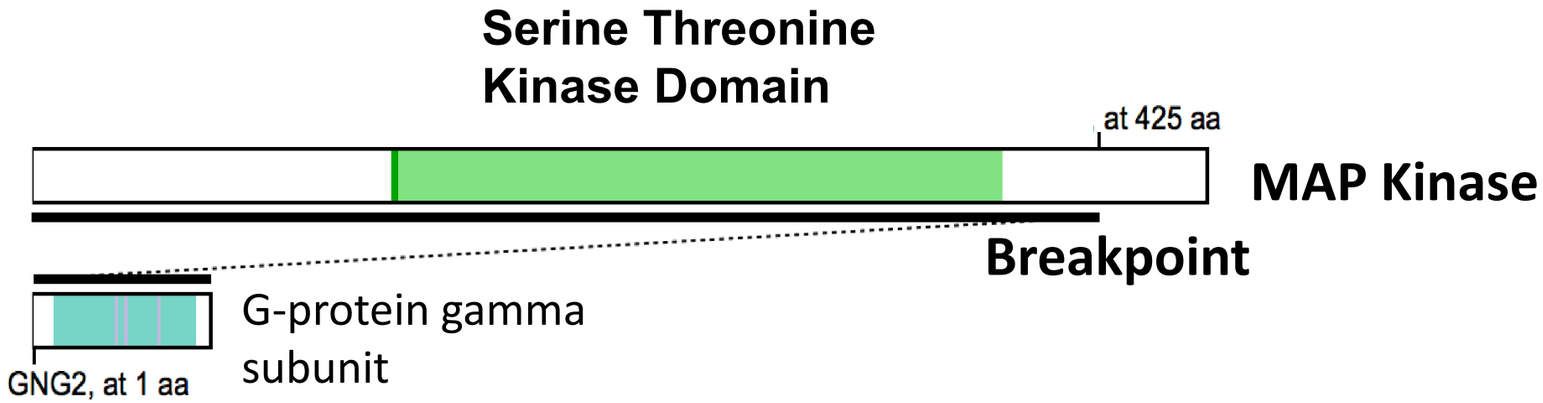
Tumor Report

Germline Report

Rusch et al, Nat. Comm. 2018

Therapy Change Based on ClinGen Data

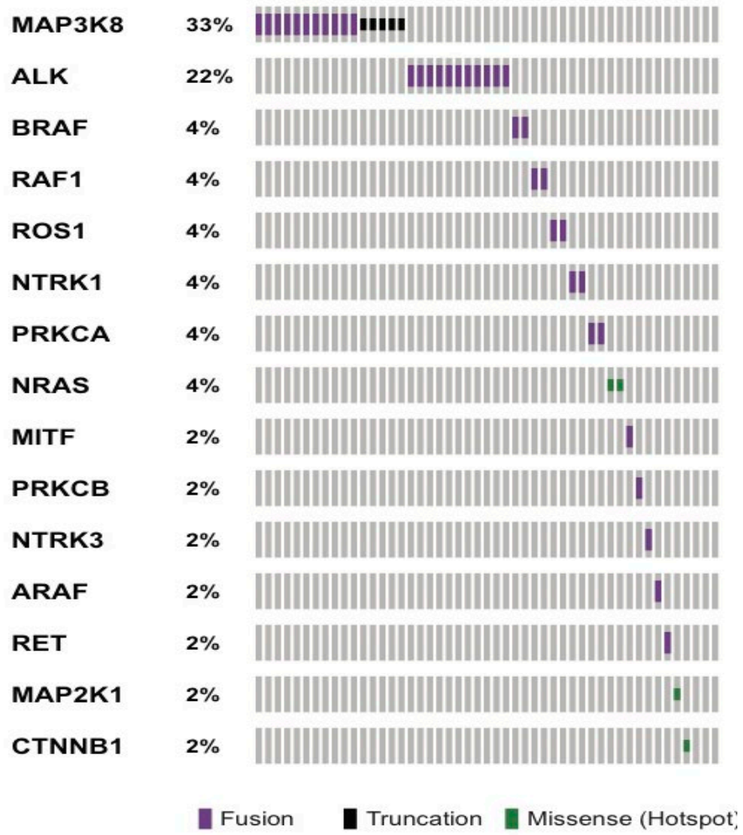
- **Child with metastatic melanoma who had failed multiple therapies**
- **Tumor analyzed by St. Jude 3-Platform Sequencing**



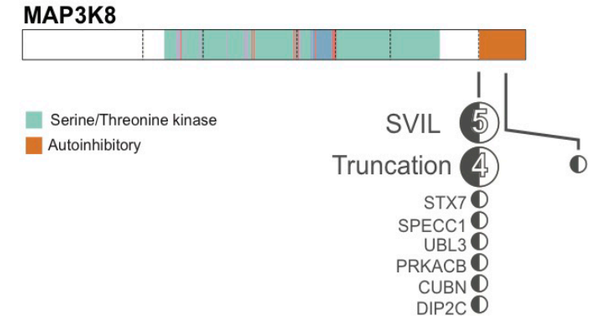
- **Activates MAP Kinase signaling independent of BRAF (unlike most melanomas)**
- **Blocked the pathway downstream of BRAF using a MEK inhibitor (trametinib) - total response but later developed resistance**

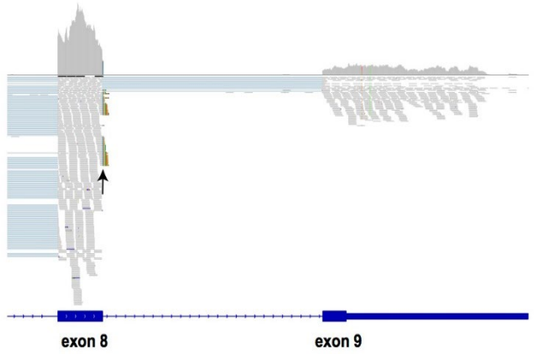
Recurrent Screening by RNA-seq of 49 FFPE Spitzoid Melanoma

MAP3K8 has the highest mutation prevalence (33%)

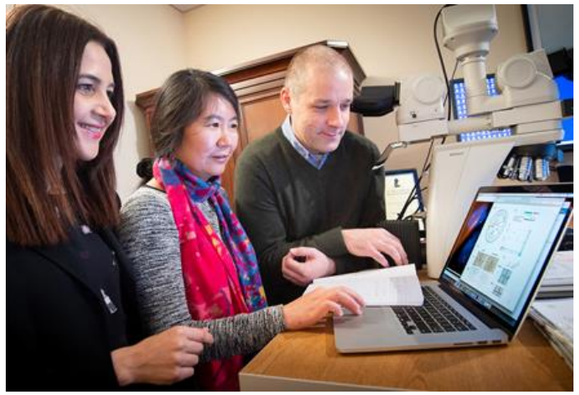


Truncations/fusions cause loss of exon 9



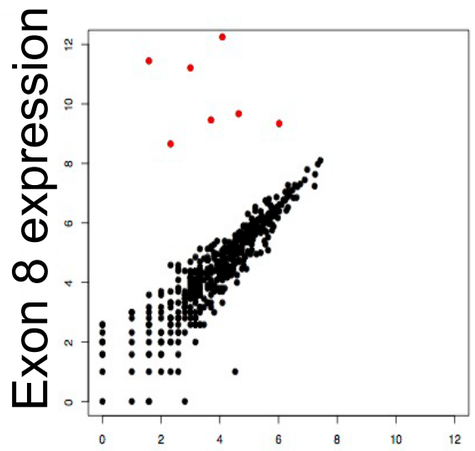


Collaboration with Richard Lee for testing new compounds targeting MAP3K8



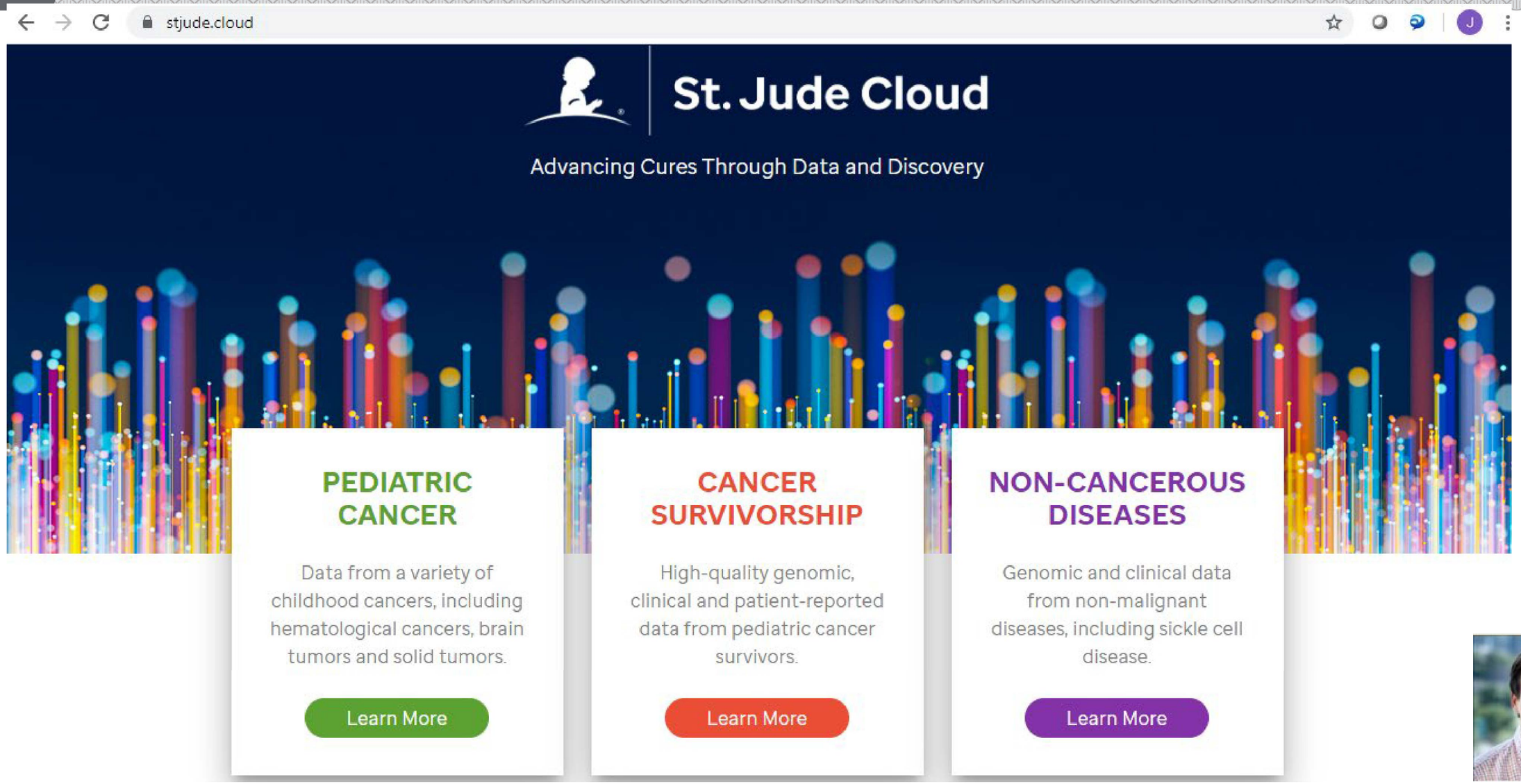
Newman *et al, Nature Medicine 2019*

472 TCGA melanoma

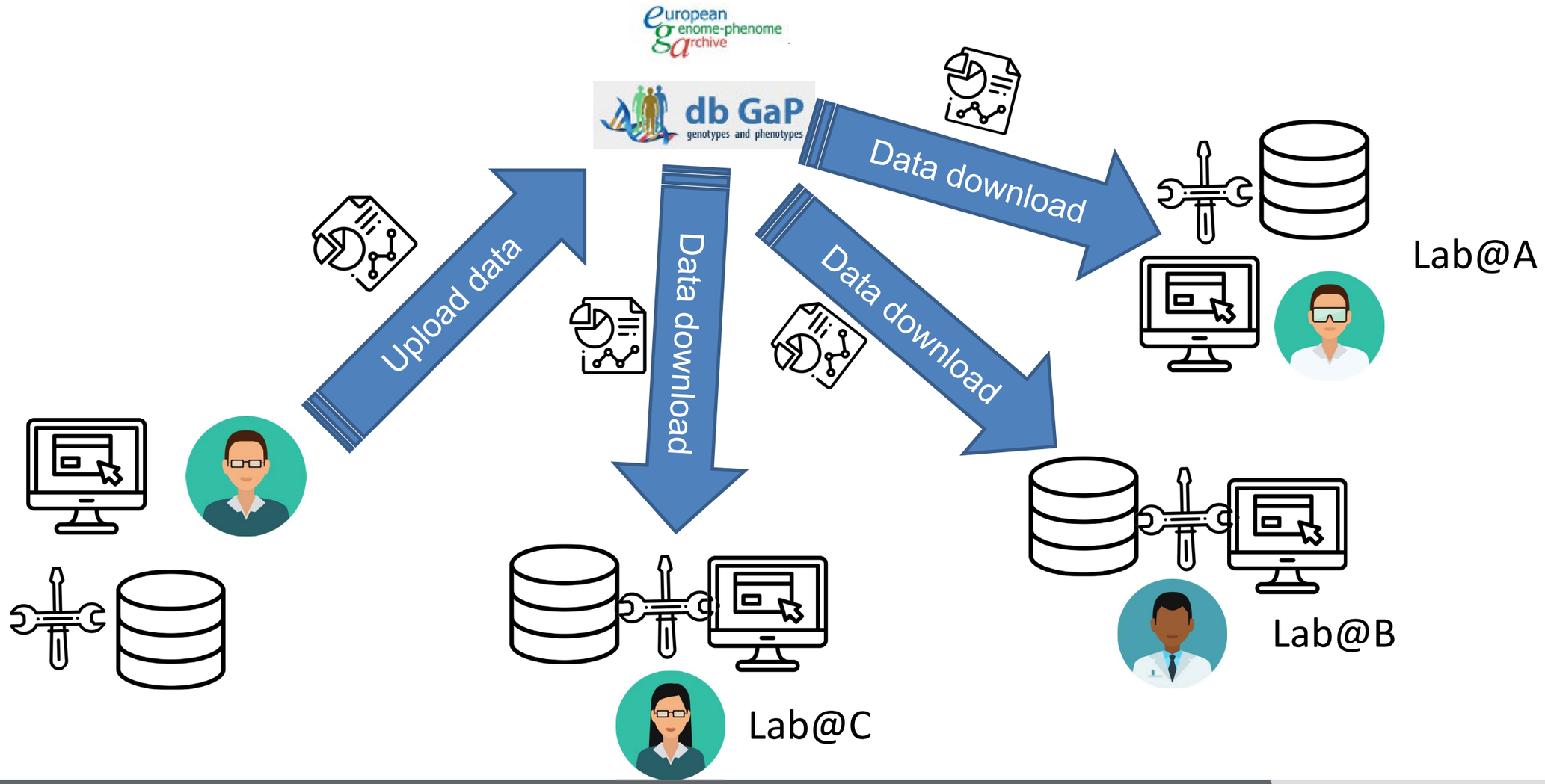


St. Jude Children's Research Hospital

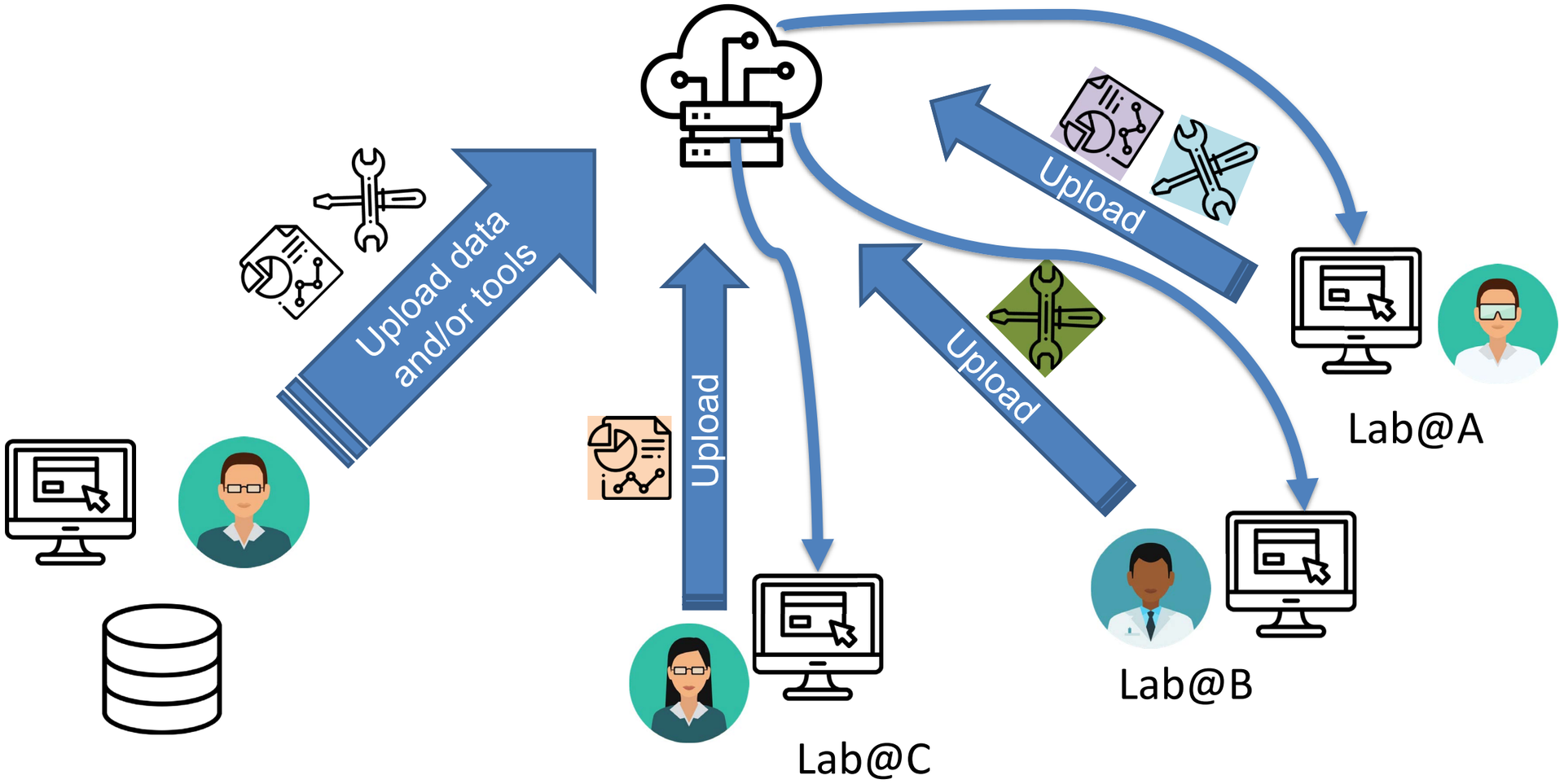
Data Sharing & Visualization on St Jude Cloud



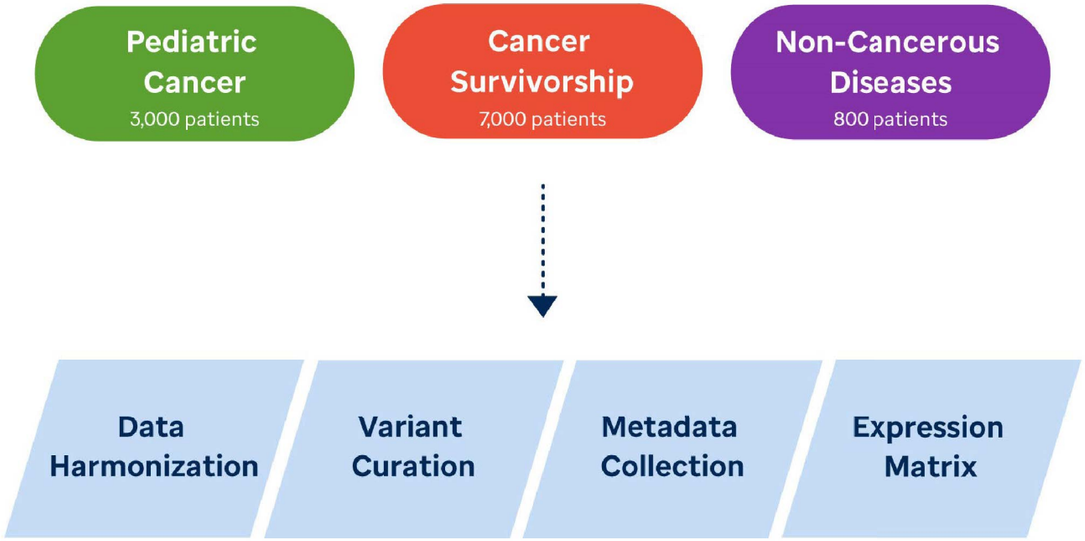
Standard Download-based Data Sharing Model

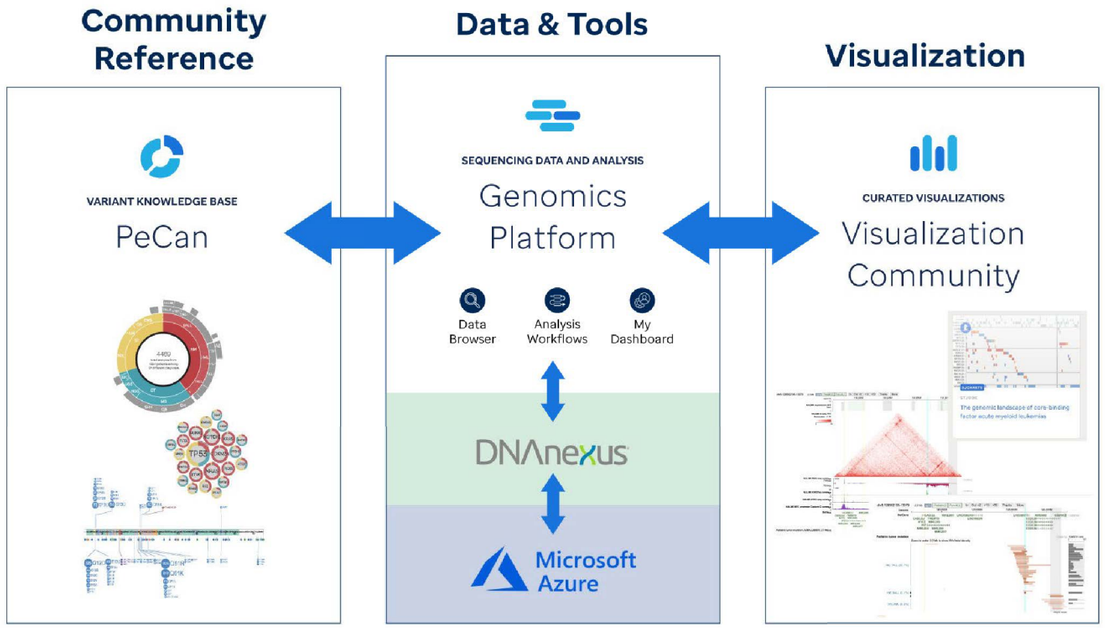


Cloud Data Sharing with Accessible Computing Infrastructure



St. Jude Cloud Ecosystem (stjude.cloud)





St. Jude Children's Research Hospital

Pediatric Cancer
Genomic Analysis
Visualization

DNAnexus
Secure cloud data host

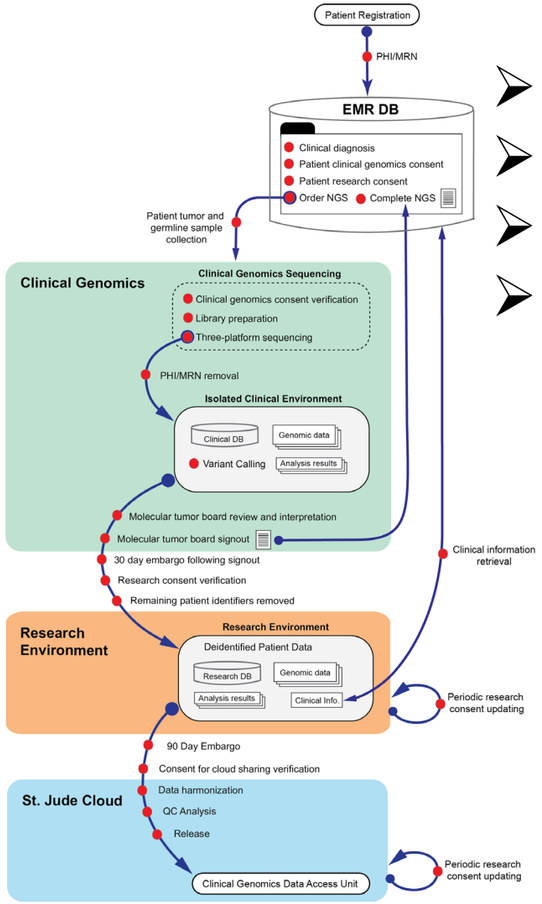
Microsoft
Azure cloud Computing

World's Largest Pediatric Cancer Genomic Data

Retrospective Research Studies		Subjects	Sequencing Type (#Samples)
PCGP	[*Tumor / Normal*]	1,610	WGS (1400), WES (1536), RNAseq (905)
Clinical Pilot	[*Tumor / Normal*]	78	WGS (156), WES (156), RNAseq (78)
St. Jude LIFE	[*Germline*]	4,833	WGS (4834), WES (3322)
CCSS	[*Germline*]	2,912	WGS (2912)
SGP	[*Germline*]	807	WGS (807)
Prospective Clinical Sequencing			
Genomes 4 Kids	[*Tumor / Normal*]	299	WGS (520), WES (559), RNAseq (254)
Clinical Genomics	[*Tumor / Normal*]	659	WGS (983), WES (1344), RNAseq (690)
Total		11,223	WGS (11,612), WES (6,917), RNAseq (1,927)

Real-time Clinical Genomics (RTCG) Streaming Enable Immediate Research

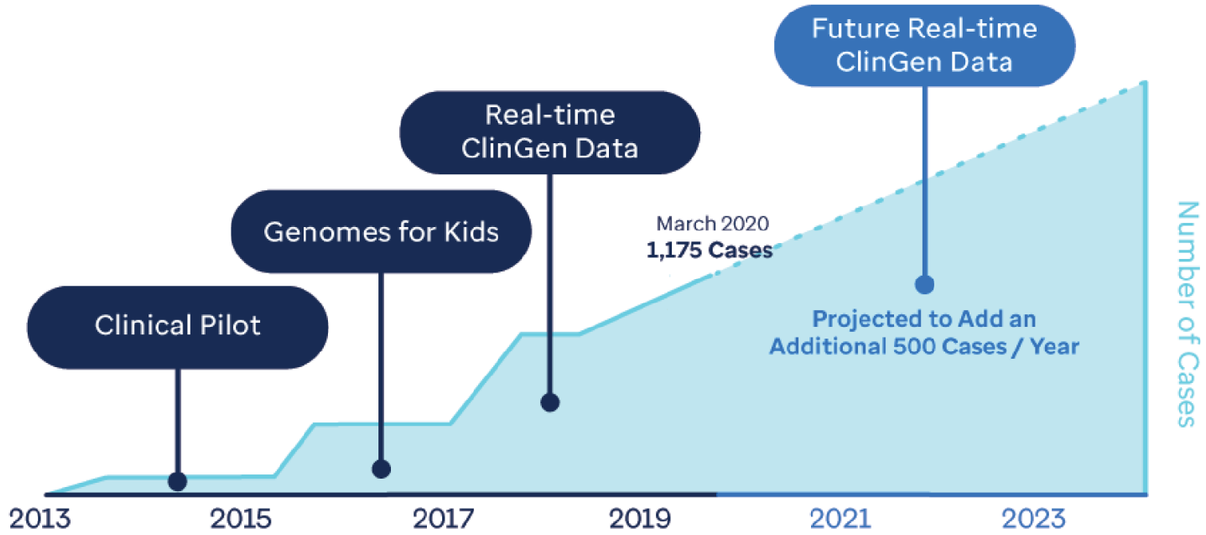
RTCG Pipeline for Regular Data Upload



- Verify Consent
- QC Check
- Meta-data Collection
- Data Harmonization

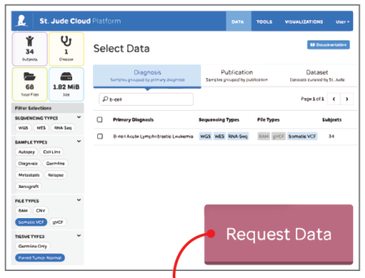
Delaram Rahbarinia

Current and Projected Data Growth

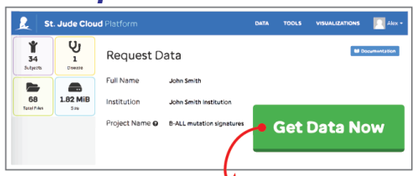


An Example of Online Analysis: Perform Mutational Signature Analysis on SJCloud

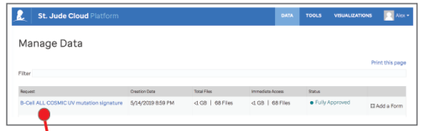
1. Select Files

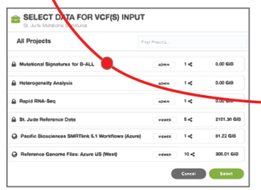


2. Request Data

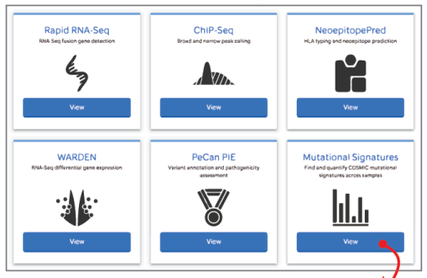


3. Create Project

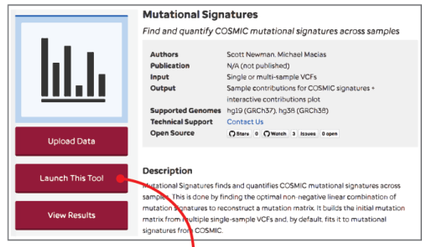




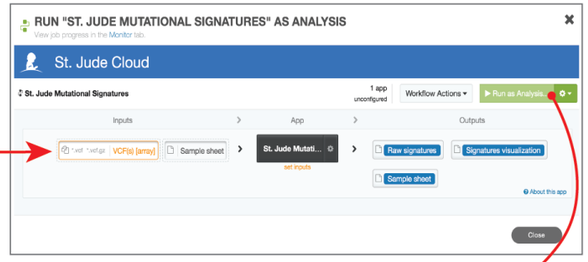
4. Select Tool

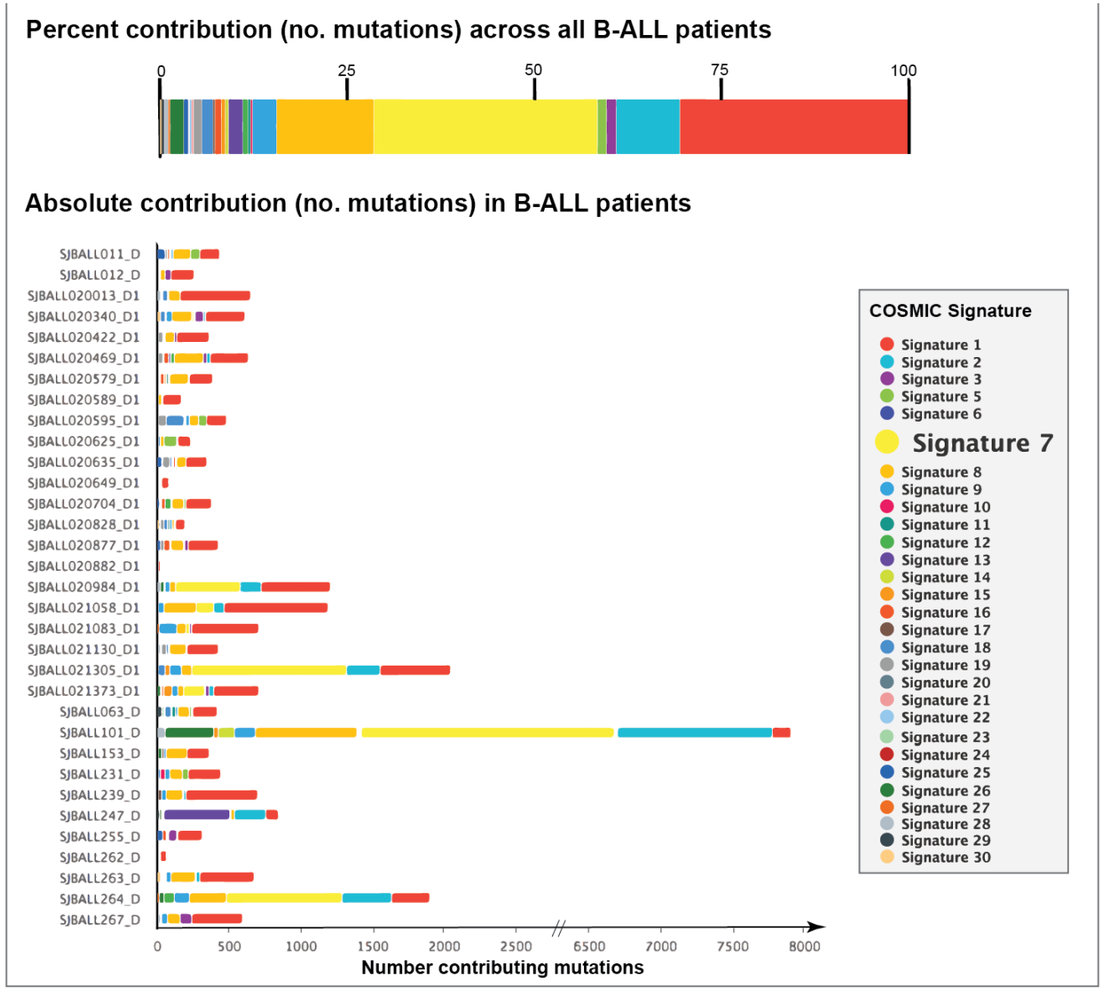


5. Launch Tool

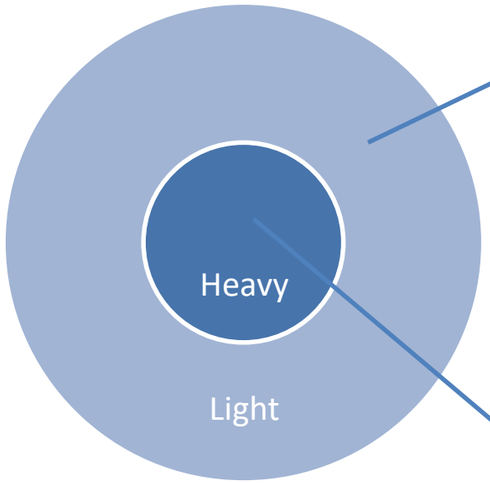


6. Set Tool Parameters





Engage St Jude Researchers During COVID-19



General Usage of St. Jude Cloud

- Carry out omics-based computation on St. Jude Cloud, removing the need for VPN or cluster access.
- New, remote-working focused quickstart guide including:
  - How to upload uploading data from cluster/laptop to cloud.
  - How to run production-grade apps at a large-scale using the cloud.
  - How to perform ad-hoc work using interactive nodes in the cloud (still in development, created in response to COVID-19).
  - How to visualize NGS data in IGV and the new GenomePaint BAM viewer (still in development, created in response to COVID-19).
- New support Slack channel: #stjudecloud-helpdesk (visit the guide on how to join the channel).
- Available to all researchers

Use general case: in an attempt to evaluate the performance of detect the subclonal SNV/Indels by using MSKCC's cfDNA data as a training set, the Ma lab uploaded the MSK data to SJCloud and was able to correct an error in the original data by curating the data via Cloud access

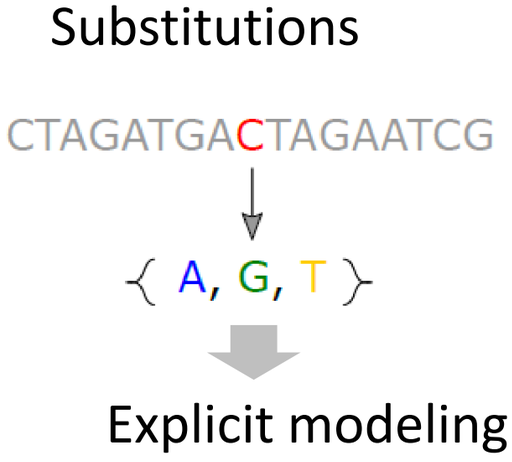
COVID-19 Discovery Program

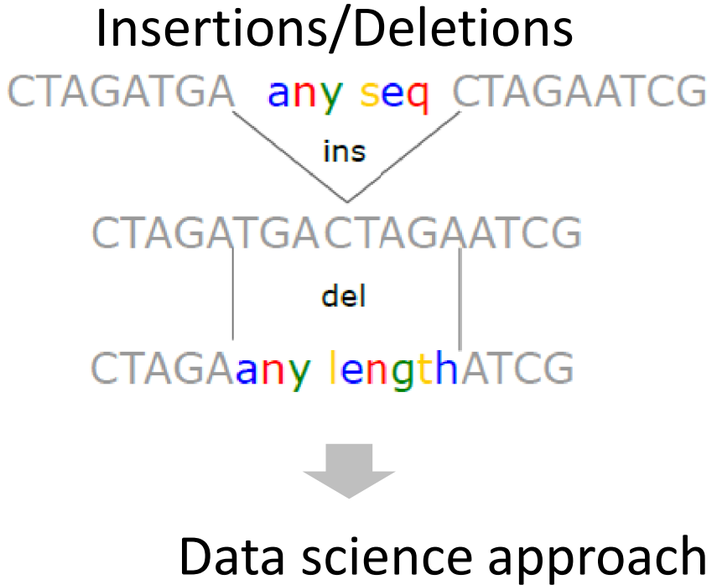
- Everything in "General usage", plus
- Sponsored compute and storage costs for your workloads in the cloud.
- Software engineering support (up to some limit that is jointly set).
- Weekly meeting with cloud team to ensure your research is moving forward effectively.
- Limited availability, by application only.

RNAIndel: An example of using SJCloud data for methods development

- **RNA-seq data are generated routinely for research and clinical testing due to low sequencing cost and data storage**
- **Expressed variants are more valuable biomarkers than DNA variant**
- **Small insertions/deletions (indels) are more challenging to model**

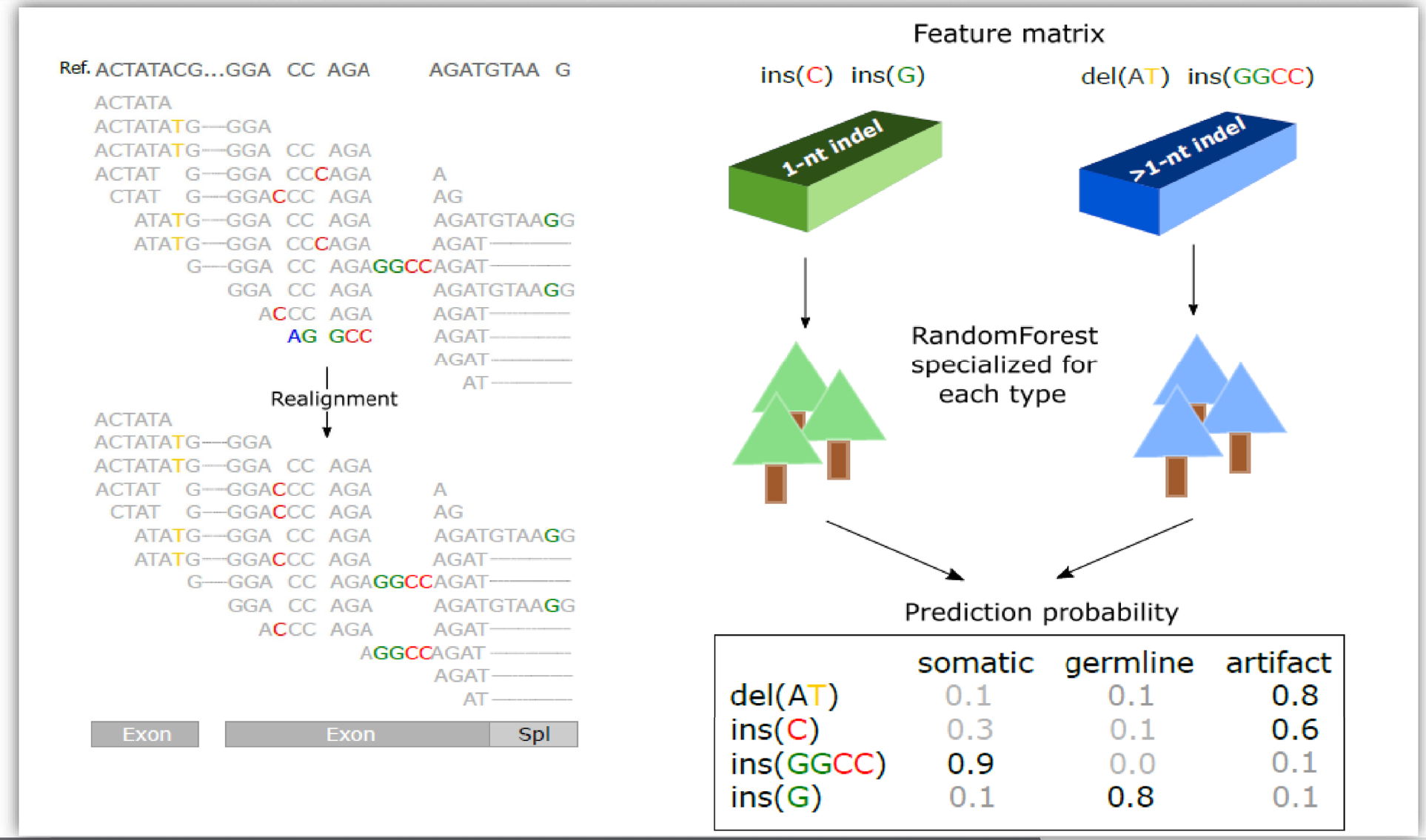
Kohei Hagiwara





RNAIndel Computation Framework

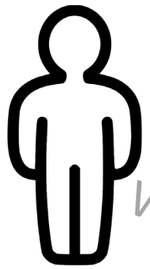
RNA.bam



	somatic	germline	artifact
del(AT)	0.1	0.1	0.8
ins(C)	0.3	0.1	0.6
ins(GGCC)	0.9	0.0	0.1
ins(G)	0.1	0.8	0.1

Constructing the Training Set using SJCloud Data

Genome4Kids



$n = 300$

RNA-Seq

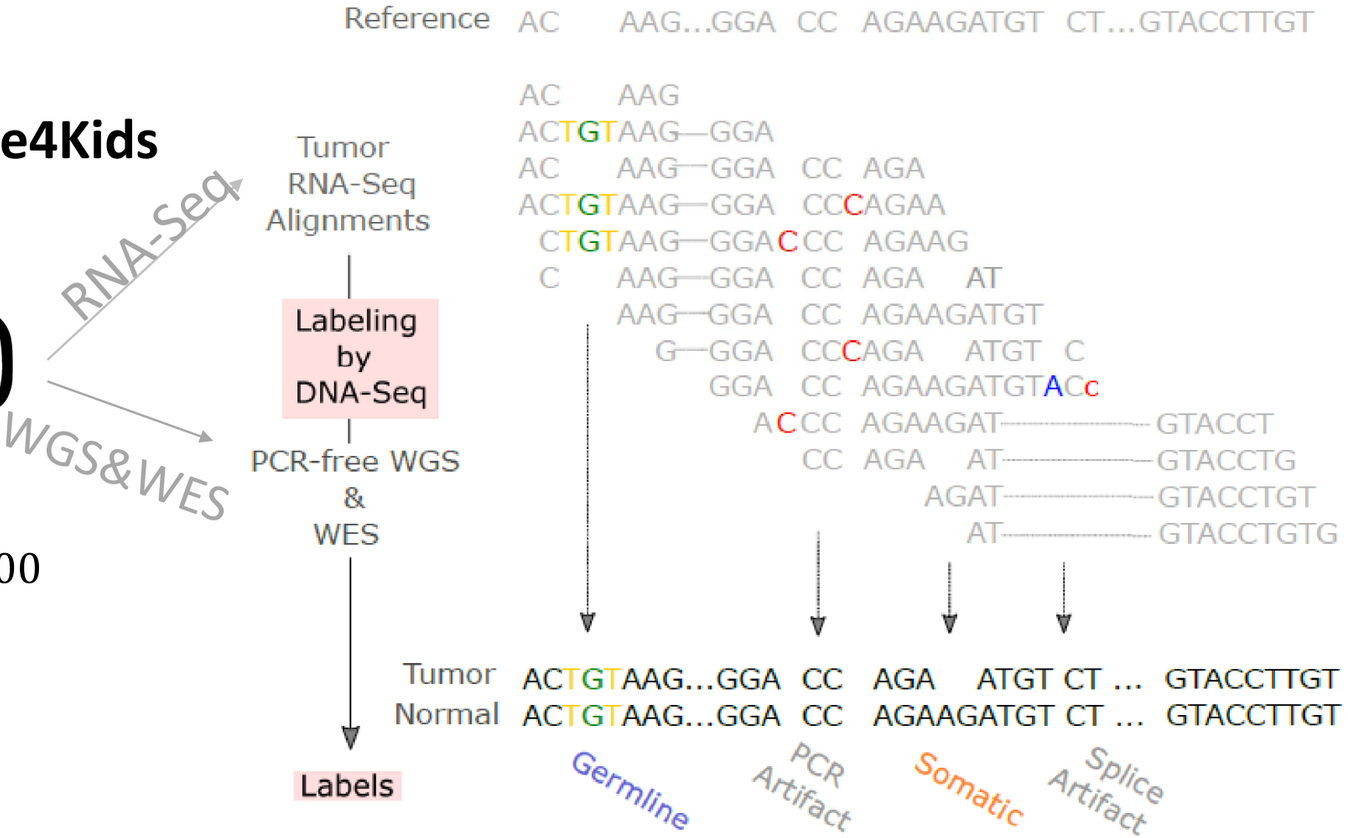
WGS&WES

Tumor RNA-Seq Alignments

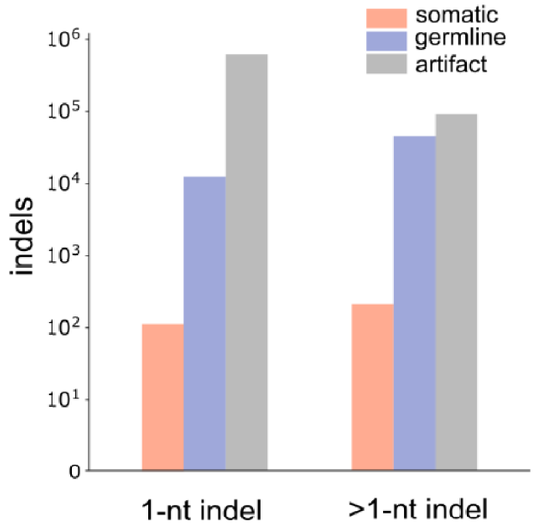
Labeling by DNA-Seq

PCR-free WGS & WES

Labels



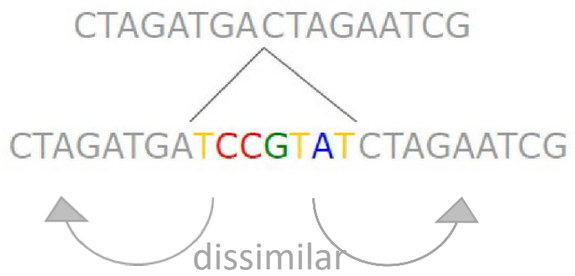
765,475 labeled RNA-Seq indels



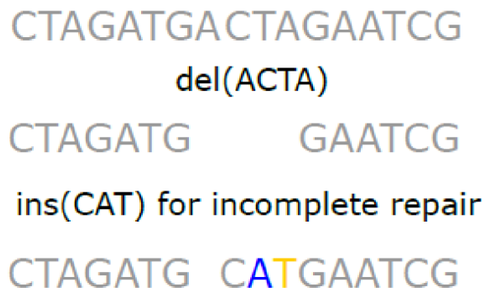
Features Distinct for Somatic and Germline Indels

Somatic indels

- Complex sequence pattern



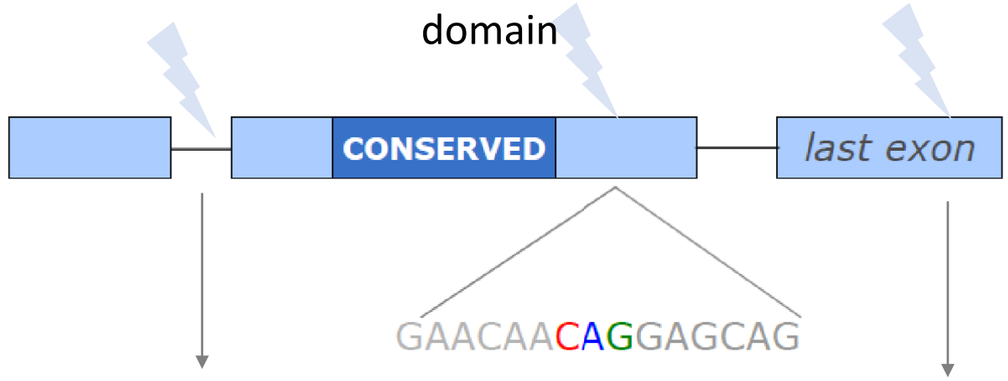
- Co-occurrence of insertion/deletion



Germline indels

Features less deleterious to the gene product

- Splice region
- Off-conserve domain
- Last exon



may Not disturb splicing

Frequently in-frame

Evade NMD (not to waste near-full length product)

NMD: nonsense-mediated decay

Performance in Pediatric and Adult Cancers

	Tumor	N	Library	ReadLen	Sequencer	SomaticIndels*		TPR	Median #FP/Sample			Median indels per sample
									A	B	C	
1	Pediatric	77	TotalRNA	100	HiSeq2000or 2500	*s*	17	0.882	3		3	2318
						m	40	0.975	4		3	311
2	AML	158	Poly-A	75	HiSeq2000	*s*	22	0.954	2		1	1036
						m	61	0.984	2		2	202
3	NSCLC	90	Poly-A	100	HiSeq1500	*s*	97	0.887	6	3	4	3171
						m	68	0.941	7	4	4	394
4	RCC	91	Poly-A	50	HiSeq2000	*s*	130	0.877	8	5	5	4303
						m	81	0.889	8	2	2	510
5	COAD (Hyper)	29	Poly-A	75	GAIIX	*s*	120	0.392	20		20	999
						m	53	0.953	11		10	141
	COAD (NonHyper)	102				*s*	30	0.9	4		3	466
						m	14	1.000	4		2	128

- Pediatric: 20 pediatric tumor types from St Jude Clinical Sequencing Pilot Study
- AML: Acute myeloid leukemia from TARGET study
- NSCLC: Non-small cell lung cancer from Nanjing study
- RCC: Renal cell carcinoma from TCGA study
- COAD: Colon adenocarcinoma from TCGA study

Discover Indels Missed by DNA Analysis

Low frequent driver mutations in highly expressed genes can be “rescued” from RNA-Seq analysis

Confirmed Indels from AML Test Data Set

ID	Gene	AAChange	VAF	500xTargetedValidation
PAKTCX	EP300	Y207fs	0.102	✓
PANLIN	CEBPA	P23fs	0.167	✓
PAPVDV	RAD21	D543fs	0.021	✓
PARSHM	KIT	Y418_D419>Y	0.148	✓
PASWPT	CREBBP	S1767fs	0.012	✓

Hagiwara et al, Bioinformatics 2020
https://github.com/stjude/RNAIndel

Future Directions

Genomic Variants

Germline Pathogenic

Non-coding Regulatory

Relapsed Cancer

Innovative Methods

Cutting-edge Analysis

Data Resources

Advance the Understanding and Treatment of Pediatric Cancer and Other Catastrophic Diseases

St. Jude Cloud

Advancing Cures Through Data and Discovery