MAYO CLINIC

CENTER FOR
INDIVIDUALIZED MEDICINE

Key Problems in Addressing the N+1 problem in Patient Care.

Daniel Quest

Note: The case study in this slide deck should not be considered clinical advice. The goal is to understand the informatics challenges!

Introduction

In 1900 David Hilbert in Paris at the International Congress of Mathematicians proposed 10 problems (later 23) that greatly influenced the progression of mathematics for the 20th century.

This is a similar journey – but in 'genomic medicine'

Today I will present 9 problems:

AND SPARK IS AWESOME AT MOST OF THEM!


4 18:48

Madelyn Shumaker

Madelyn was 8 years old when diagnosed with DIPG (diffuse intrinsic pontine glioma), which is nearly always fatal and lacks an effective treatment. Nearly all die within two years. She underwent 'personalized medicine' in an attempt to target her cancer.

January 29, 2015: Maddie goes to St. Jude Children's hospital for 6 weeks chemo/radiation as part of a DIPG clinical trial.

January 26, 2015: Maddie is diagnosed with DIPG

March 30, 2015: Maddie returns home and starts school again. Maddie, has some slight hearing loss

April 15, 2015: Maddie returns to St. Jude's for a follow up MRI. The results show the cancer virtually gone from the brainstem

May 22, 2015: Maddie undergoes a second round of chemo

June 12, 2015: Maddie returns to St. Jude's for a follow up MRI. There is evidence of necrosis (cell death) in the brainstem.

October 20, 2015: Maddie's symptoms worsen and become persistent, the cancer is recurrent

October 29, 2015: Maddie's undergoes a revolutionary surgery at Sloan Kettering where they do a biopsy on the tumor for molecular and pathology analysis

December 1, 2015: Maddie goes begins to have adverse reactions to chemo

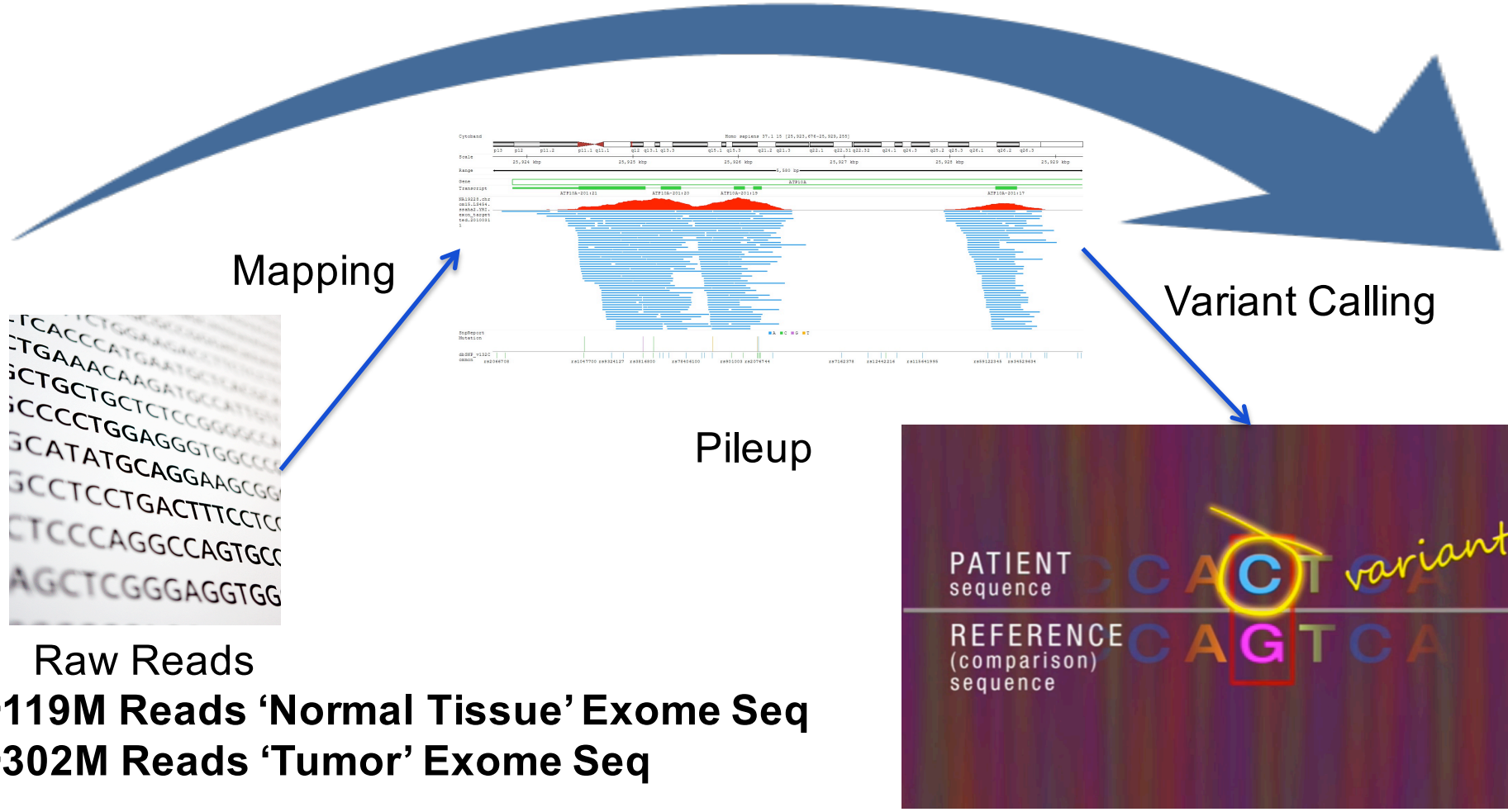
November 14, 2015: Maddie goes to Dr. Giselle Sholler Helen DeVos Children's Hospital in Grand Rapids, Mich for genome and transcriptome sequencing. The Tumor board recommends treatment based on the findings

December 10, 2015: Maddie passes away from DIPG

Data Types (Available to the Tumor Board)

- NGS Whole Exome (Somatic and Germline)
- RNAseq
- PDFs Describing Drug/Gene Interactions

Problem 1: Variant Calling



Raw Reads

~119M Reads 'Normal Tissue' Exome Seq

~302M Reads 'Tumor' Exome Seq

https://github.com/bigdatagenomics/ - Variant Calling in Spark!

Genomic Variants

62,116 Variants'Normal Tissue'

4361 Variants 'Tumor'

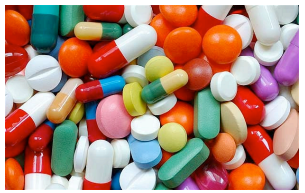
Problem 2: Annotation

http://bioinformaticstools.mayo.edu/research/bior/

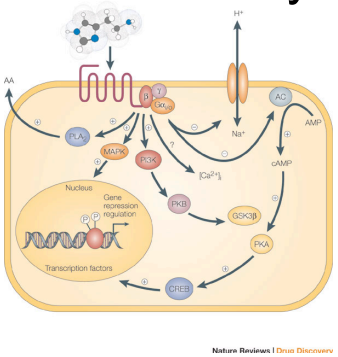
Problem 2: Annotation

Information from data sources from other organizations and institutions that give important and actionable background.

Drugs



Gene Functions and Pathways



Our Data:
Genome, Transcriptome, Epigenome, Microbiome, Proteome

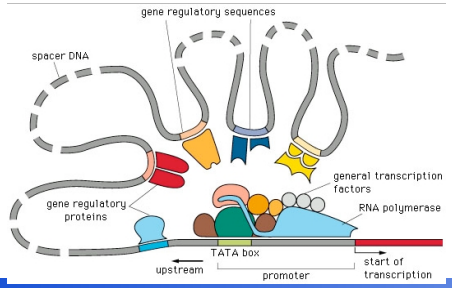
Mayo Clinical Knowledge

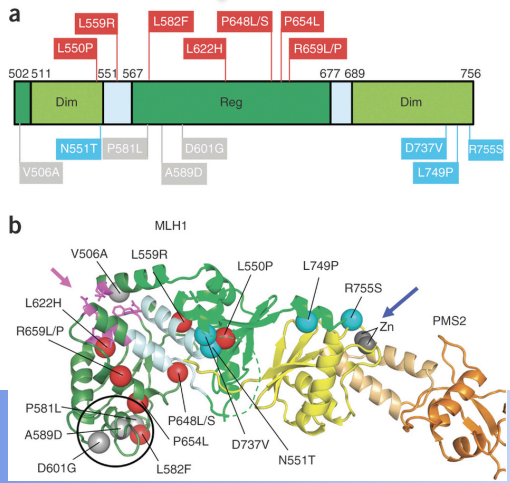
MAYO CLINIC

Oncogenes, tumor suppressors, epigenetic readers/writers, etc.

Population Variation

Gene regulation





Functional Impact

Document Data Model

Original Data (normalized syntax)

Added Data
(normalized syntax + semantics)

CATALOG

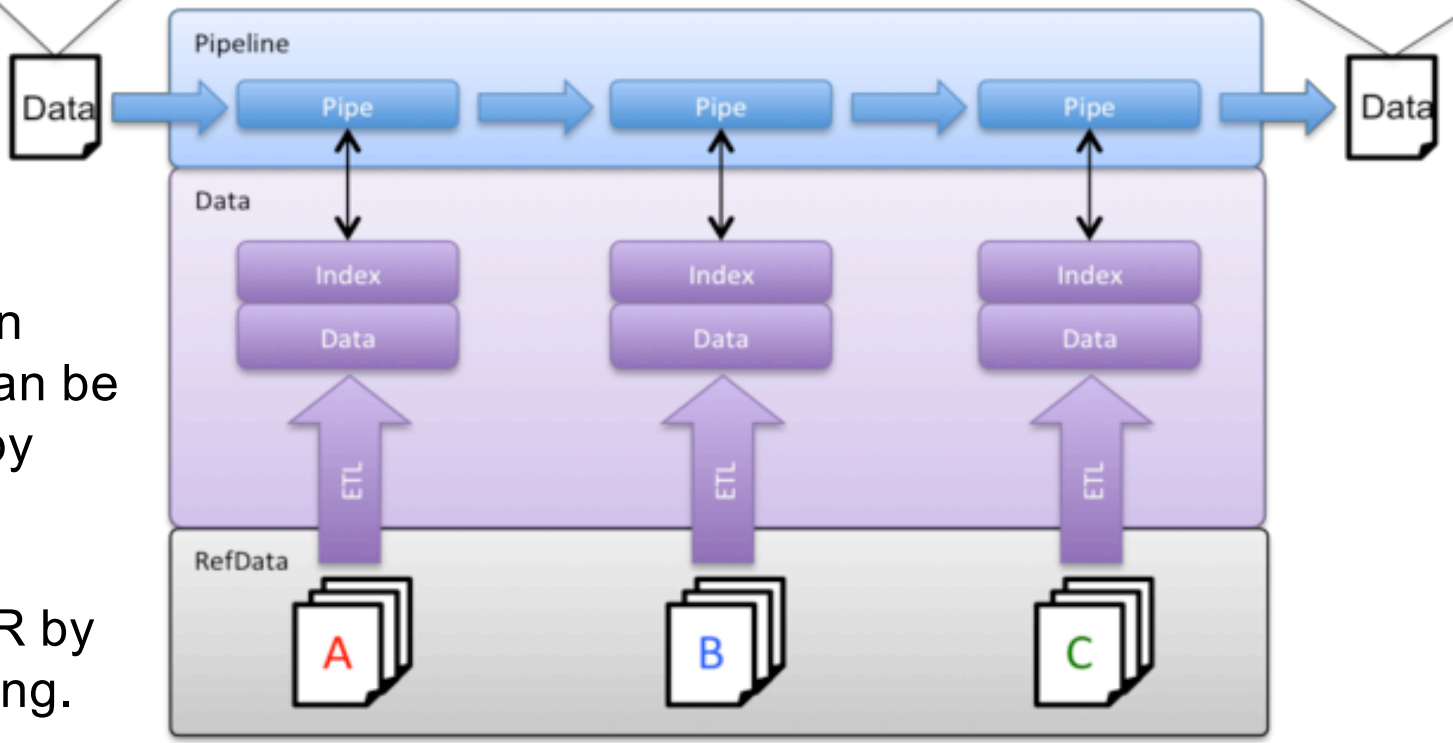
Below is the corresponding Catalog structure for variant **rs10399749**.


```
{
    "CHROM": "1",
    "POS": "55299",
    "ID": "rs10399749",
    "REF": "C",
    "ALT": "T",
    "QUAL": ".",
    "FILTER": ".",
    "INFO": {
        "RSPOS": 55299,
        "GMAF": 0.2537,
        "dbSNPBuildID": 119,
        "SSR": 0,
        "SAO": 0,
        "VP": "0501000000050301170000100",
        "WGT": 1,
        "VC": "SNV",
        "SLO": true,
        "ASP": true,
        "G5A": true,
        "G5": true,
        "GNO": true,
        "KGPhase1": true,
        "KGPROD": true,
        "OTHERKG": true,
        "PH3": true
    },
    "_id": "rs10399749",
    "_type": "variant",
    "_landmark": "1",
    "_refAllele": "C",
    "_altAlleles": [
        "T"
    ],
    "_minBP": 55299,
    "_maxBP": 55299
}
```


BioR Annotation Engine

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr2	48032098	.	A	T	.	PASS	DP=100
chr2	220462640	.	G	T	.	PASS	DP=100
chr4	54417522	.	A	G	.	PASS	DP=100
chr5	79950733	.	C	G	.	PASS	DP=100

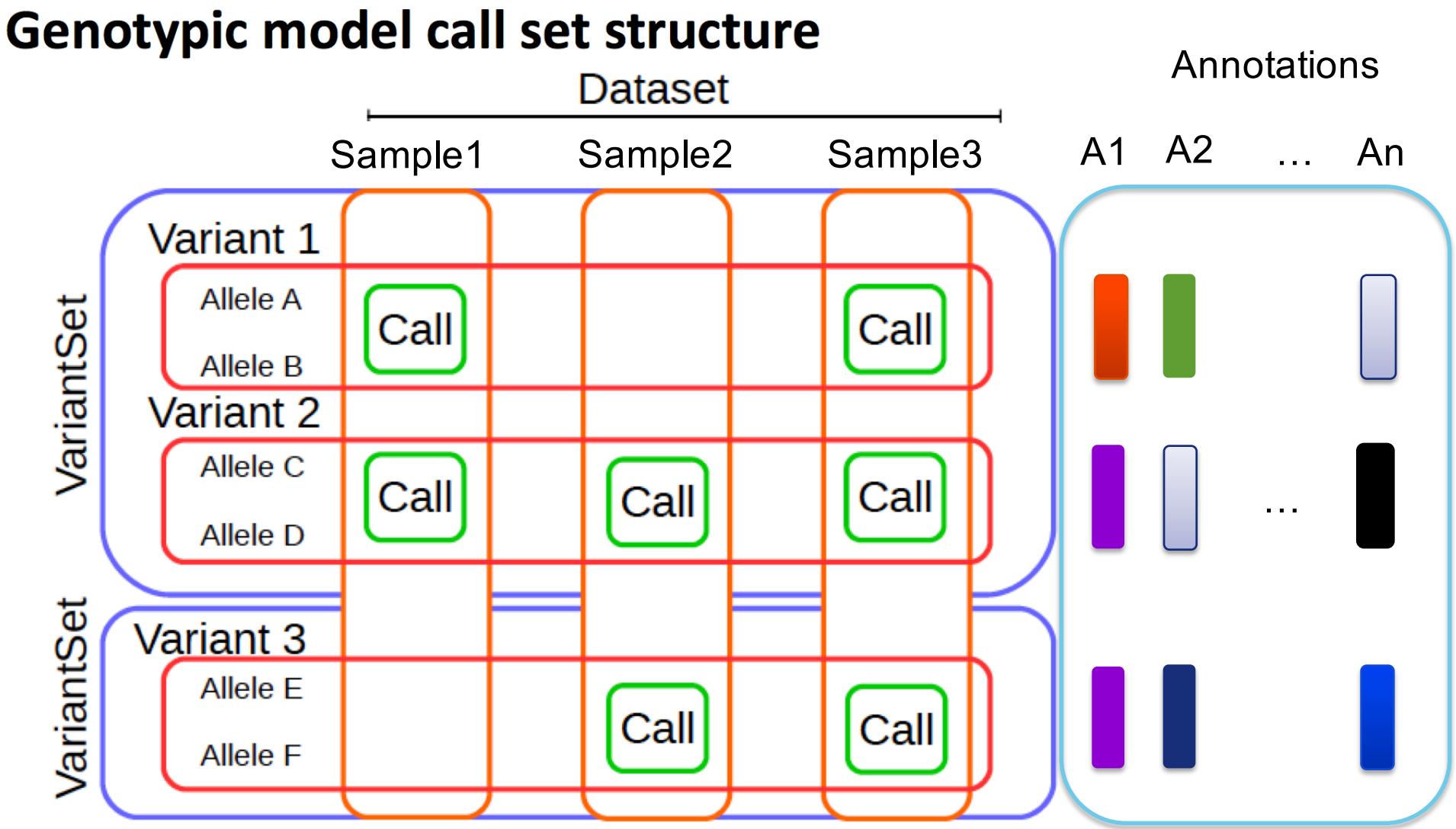
CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	A	B	C
chr2	48032098	.	A	T	.	PASS	DP=100	present	tolerated	0.001
chr2	220462640	.	G	T	.	PASS	DP=100	present	tolerated	0.239
chr4	54417522	.	A	G	.	PASS	DP=100	absent	tolerated	0.05
chr5	79950733	.	C	G	.	PASS	DP=100	pressent	damaging	1.009



>90% of annotation queries can be handled by genomic position search OR by ID matching.

This is a Map-Reduce Problem! Spark to the rescue!

Problem 3: Variant Filtering



Problem 3: Variant Filtering

Genetic Councilors at Mayo are Using VCF Miner in the Clinic to find the cause of disease

Each case requires a different ‘schema’ because each disease is different.

http://bioinformaticstools.mayo.edu/research/vcf-miner/

VCF-Miner Stats for Madelyn

#	Filter	Count
1	All Variants	61971 (whole exome – whole genome is usually ~3-5M)
2	Germline Mutations (Maddie relative to HG19)	54579
3	Somatic Mutations (tumor only)	7392
4	Variants in Cancer Genes (cosmic 595) && 3	202
5	SNPEFF Impact = HIGH || MODERATE && 3	257
6	Polyphen = possibly damaging, probably damaging, unknown && 3	93
7	SIFT_TERM && 3	69
8	4 && 5	8
9	Variants Filtered by Annovar	83
10	Variants in final report to Tumor Board	7 (1 RARG, 6 PIK3CA)

VCF-Miner | Home | Advanced | shumakerannotated.results.vcf.gz | Daniel Quest | Logout

Variants | Columns | Export

Show Analysis

25 records per page

Showing 1 to 8 of 8 entries

CHROM	POS	ID	REF	ALT	#_Samples	Samples	SNPEFF_Effect	SNPEFF_Gene_name
1	226252135	.	A	T	1	MGT9-209-08_EXO_T	NON_SYNONYMOUS_CODING	H3F3A
2	158630626	.	C	T	1	MGT9-209-08_EXO_T	NON_SYNONYMOUS_CODING	ACVR1
3	178936091	.	G	A	1	MGT9-209-08_EXO_T	NON_SYNONYMOUS_CODING	PIK3CA
4	1809110	.	CTG	C	1	MGT9-209-08_EXO_T	FRAME_SHIFT	FGFR3
4	54319247	.	CAG	C	1	MGT9-209-08_EXO_T	FRAME_SHIFT	FIP1L1
6	29911901	.	C	G	1	MGT9-209-08_EXO_T	NON_SYNONYMOUS_CODING	HLA-A
6	29911970	.	G	A	1	MGT9-209-08_EXO_T	NON_SYNONYMOUS_CODING	HLA-A
X	123224754	.	A	T	1	MGT9-209-08_EXO_T	NON_SYNONYMOUS_CODING	STAG2

← Previous 1 Next →

- In Clinical Report
- In Raw Report
- In DIPG Literature – not in report
- Novel found by VCF-Miner!
  Missing in VCF-Miner 'Damaging' Analysis - **RARG** (there are two if we consider all somatic variants)

Problem 4 Clinical Oncology

Mutations in Genes Relevant to Cancer*

Gene name	Mut type	Location	Specific change	COSMIC	Gene description (NCBI)
ACVR1 (activin A receptor type I)	SNV	chr2:158630626, Pfam Domain: Transforming growth factor beta type I GS-motif	Missense, R206H, DNA allele ratio 0.46	150 coding mutations, 12 R206H	Activins are dimeric growth and differentiation factors which belong to the transforming growth factor-beta (TGF-beta) superfamily of structurally related signaling proteins. Activins signal through a heteromeric complex of receptor serine kinases which include at least two type I (I and IB) and two type II (II and IIB) receptors. These receptors are all transmembrane proteins, composed of a ligand-binding extracellular domain with cysteine-rich region, a transmembrane domain, and a cytoplasmic domain with predicted serine/threonine specificity. Type I receptors are essential for signaling; and type II receptors are required for binding ligands and for expression of type I receptors. Type I and II receptors form a stable complex after ligand binding, resulting in phosphorylation of type I receptors by type II receptors. This gene encodes activin A type I receptor which signals a particular transcriptional response in concert with activin type II receptors. Mutations in this gene are associated with fibrodysplasia ossificans progressive.
PIK3CA (phosphatidylinositol-4,5,-bisphosphate 3-kinase, catalytic subunit alpha)	SNV	chr3:178936091, Pfam Domain: Phosphoinositide 3-kinase family, accessory domain (PIK)	Missense, E545K, DNA allele ratio 0.14	8178 coding mutations, 1277 E545K	Phosphatidylinositol 3-kinase is composed of an 85 kDa regulatory subunit and a 110 kDa catalytic subunit. The protein encoded by this gene represents the catalytic subunit, which uses ATP to phosphorylate PtdIns, PtdIns4P and PtdIns(4,5)P2. This gene has been found to be oncogenic and has been implicated in cervical cancers.

Molecular Guided Report

Variant Type: Known Variants

Gene	AA Change	Genomic Event	Drug / Indication
PIK3CA	E545K	SNV	sirolimus, temsirolimus, everolimus / Sensitive
PIK3CA	H1047R	SNV	sirolimus, temsirolimus, everolimus / Sensitive
PIK3CA	E545K	SNV	Erlotinib/Gefitinib / Resistant
PIK3CA	H1047R	SNV	Erlotinib/Gefitinib / Resistant
PIK3CA	E545K	SNV	Imatinib / Resistant
PIK3CA	H1047R	SNV	Imatinib / Resistant

Variant Type: Variants of Unknown Significance

Gene	AA Change	Genomic Event	Drug / Indication
RARG	D95G	SNV	Retinoic Acid / Sensitive

NMTRC 009 Treatment Memo

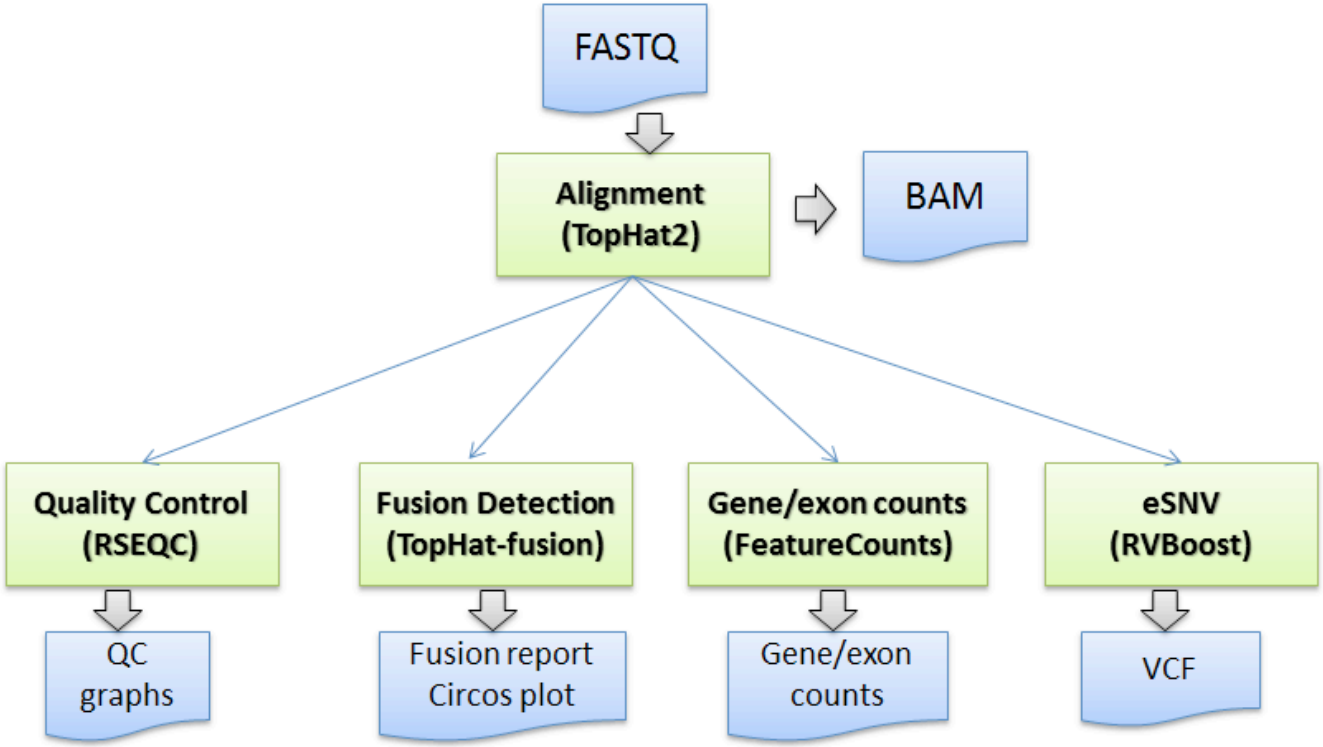
Molecular Tumor Board held on : 11 / 16 / 2015 **Subject Study ID:** MGT9-209-08

Chemotherapy Administration:

Cycles will be 21 day cycles of:

	Drug Name	Dose	Route	Schedule
1	**Etoposide**	125 mg/m2/dose	IV	Give on Days 1-3 of each 21 day cycle.
2	**Dasatinib**	65 mg/m2/dose	PO	Take 50mg orally twice daily on every day of a 21 day cycle.
3	**Temsirolimus**	35 mg/m2/dose	IV	Give on Days 1, 8, and 15 of a 21 day cycle. Pre-medication with Benadryl 30 minutes prior to dose.
4	**Vandetanib**	65 mg/m2/day	PO	Take 50mg (1/2 of 100mg tablet) once daily on every day of a 21 day cycle. *Please ask pharmacy to cut tablets in half prior to dispensing May replace with Thalidamide if not able to order Vandetanib due to limited access program.

Problem 5: Integrate RNA and other data to get improved accuracy.

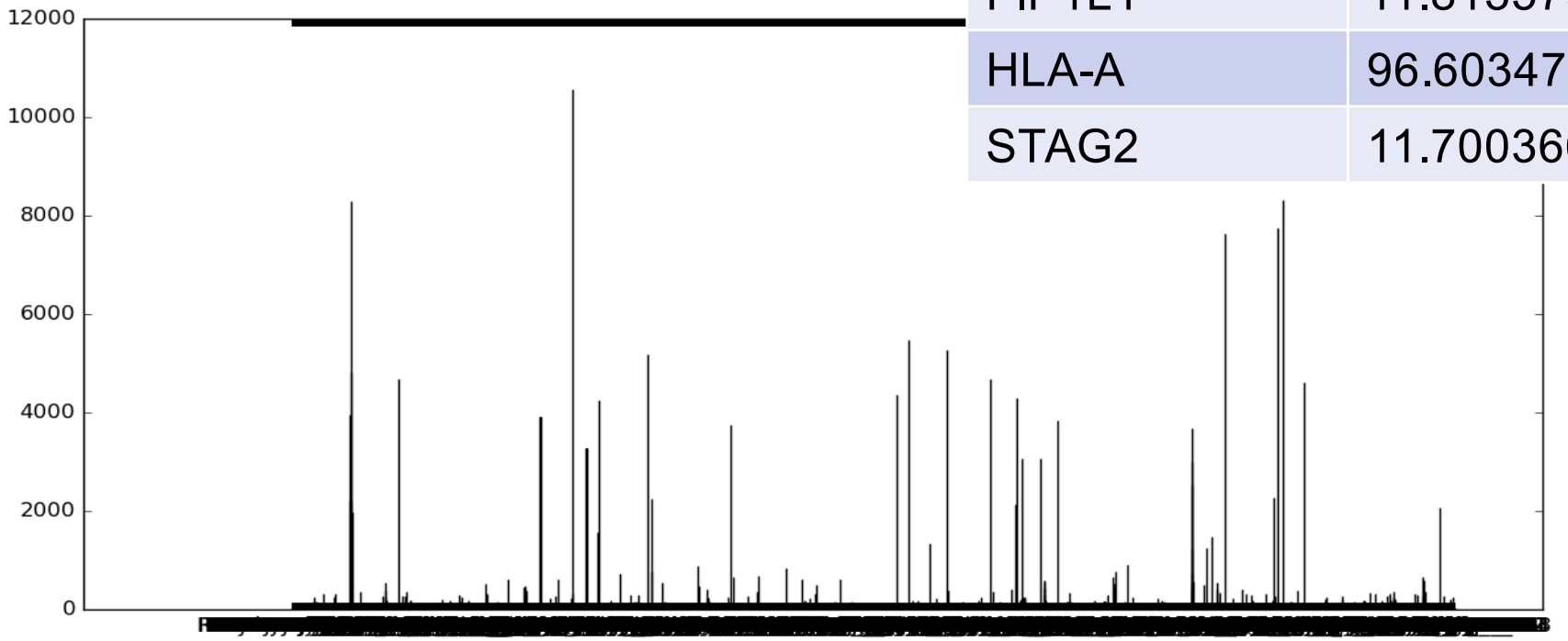


MAPRSeq

RNAseq Stats:

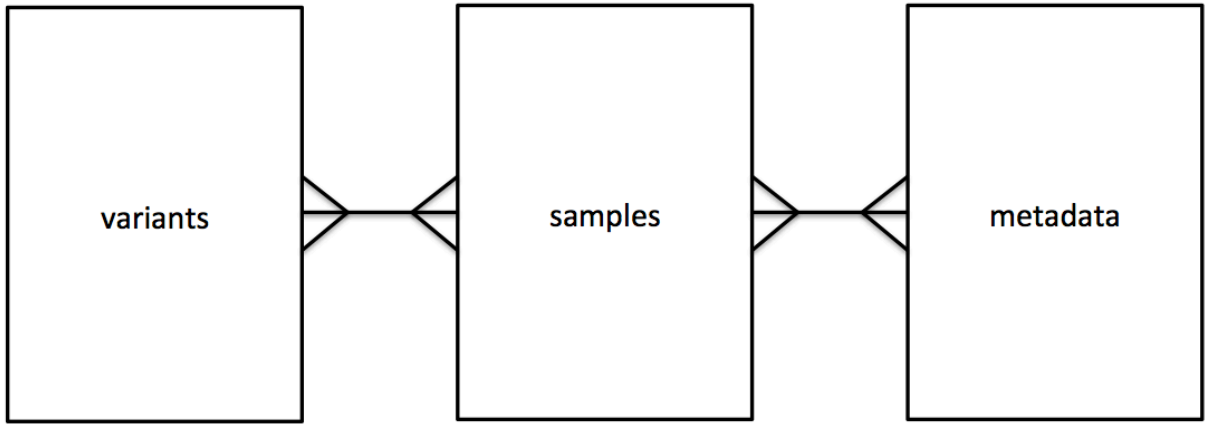
	Number of Transcripts
RPKM > 1	17651
RPKM > 0	38320
ALL	57773

Gene	RPKM
Average RPKM >1	16.2921
H3F3A	23.844392374
ACVR1	7.87217392506
PIK3CA	3.67493145428
FGFR3	27.3744042904
FIP1L1	11.8135796235
HLA-A	96.6034792985
STAG2	11.7003663476



Problem 6: Metadata

Samples and the metadata about them link genotypes and phenotypes



- Studies on the research side mostly 'managed' in excel files – semi-structured/denormalized.
- Each investigator collects the information they need to answer a specific question.
- **Can come from clinical notes in the EMR; free text – this requires NLP!**
- Limited information can come from the EDT for example ICD 10 codes, birthdate, ect.
- Information is decentralized and not easy to query

Diagnosis	UC vs. IC Vs. CD
Date of Current Diagnosis	date listed in Month/Year format
Initial Diagnosis	for example if 1st dx as UC, then dx with CD
Date of Past Diagnosis	date listed in Month/Year format
Gender	Male / Female

Birth date	month/day/year
Race	
Ethnicity	

Pyoderma Gangrenosum	Yes or No
Erythema Nodosum	Yes or No
Metastatic Crohn's disease	Yes or No
Uveitis/iritis	Yes or No
Episcleritis/scleritis	Yes or No
Primary Sclerosing Cholangitis	Yes or No
Arthritis - small joints (hot swollen joints)	Yes or No
Arthritis - large joints (hot swollen joints)	Yes or No
Amyloidosis	Yes or No
Ankylosing Spondylitis	Yes or No
Sacroiliitis	Yes or No
IBD-related mouth ulcers	Yes or No
Venous thrombosis	Yes or No
Arterial thrombosis	Yes or No
Kidney stones	Yes or No
Colon or rectal cancer	Yes or No
perianal procedure	Yes or No
abcess	Yes or No
fistula	Yes or No
stricture	Yes or No
seton placed	Yes or No

Crohn Phenotype	B1 or B2 or B3
Crohn's Location	Ileocolonic, Colonic, Ileum

UC Location	E1 or E2 or E3

Number of IBD related surgeries	# value
Number of resections	# value
Number of stricturoplasties	# value
Anti-TNF Ever	Yes or No
Any 1st degree family members with CD	Yes or No
Any 1st degree family members with UC	Yes or No
Ever Smoked	Yes or No
Smoke When Diagnosed	Yes or No
Currently Smoke	Yes or No
Current Smoke Amount	packs per day, can be decimal, or pack per week
Average Packs Per Day Over History	packs per day, can be decimal, or pack per week
Years Smoked	# value

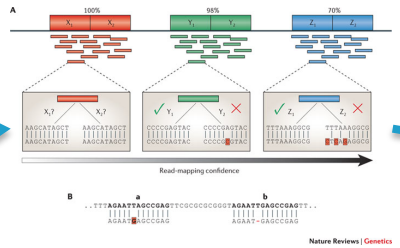
Medication for IBD	See separate list

Example Study Collection Form

Review: Components

Other Pipelines:
- MAP-RSEQ
- CHIP-SEQ
- Methyl-SEQ
- Microbiome
- Biomarkers

Sequencing



Primary Analysis
(read mapping)

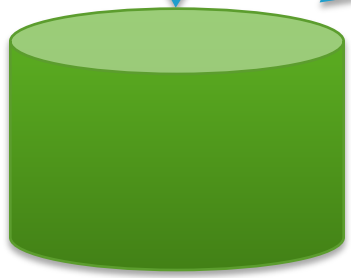
GenomeGPS
Variant
Calling

Metadata

VCF-Miner
Filtering
Sorting

BioR

VCF-Export



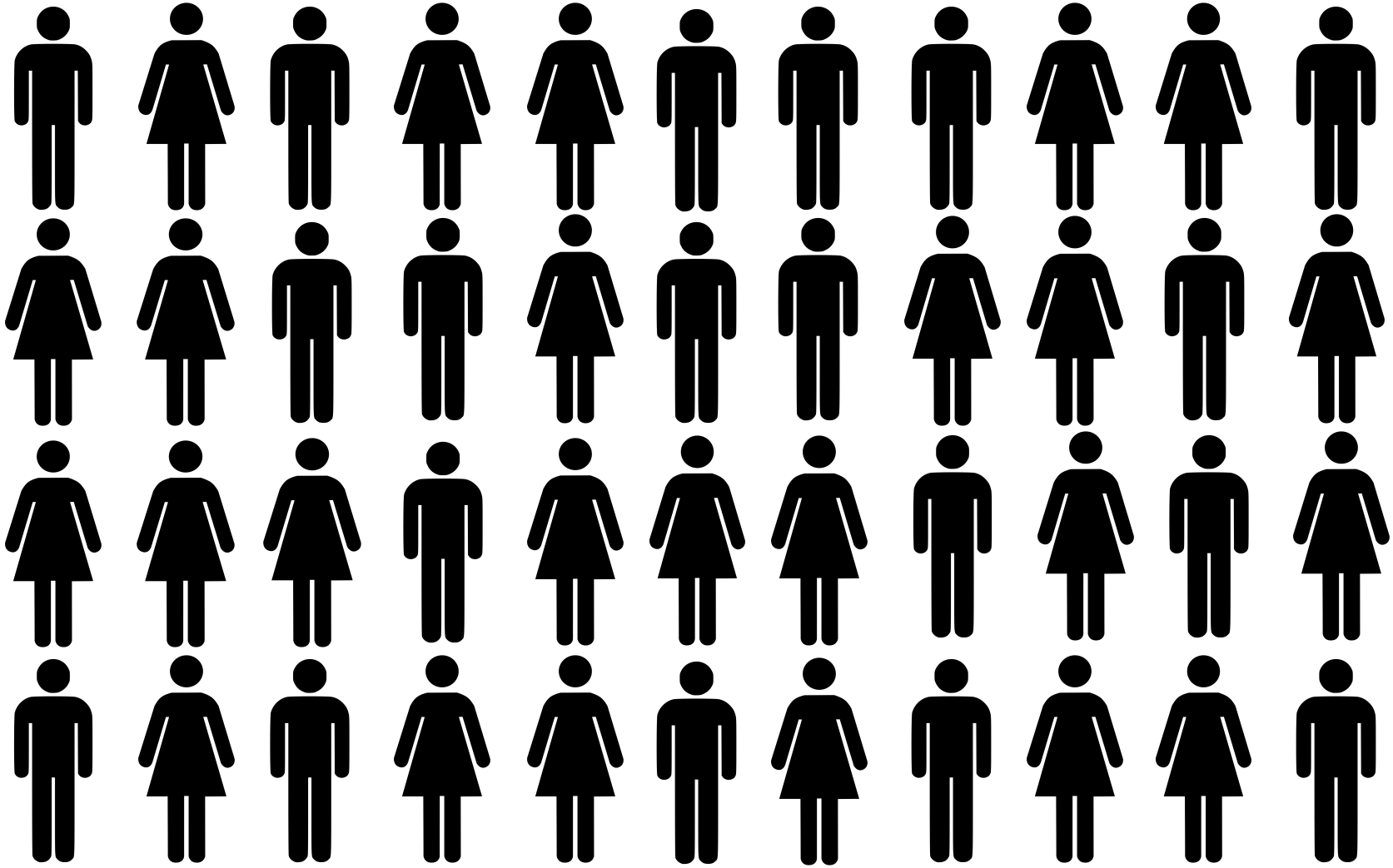
Variant
Warehouse

CUSTOM
Linux
Tooling / R

Annotation

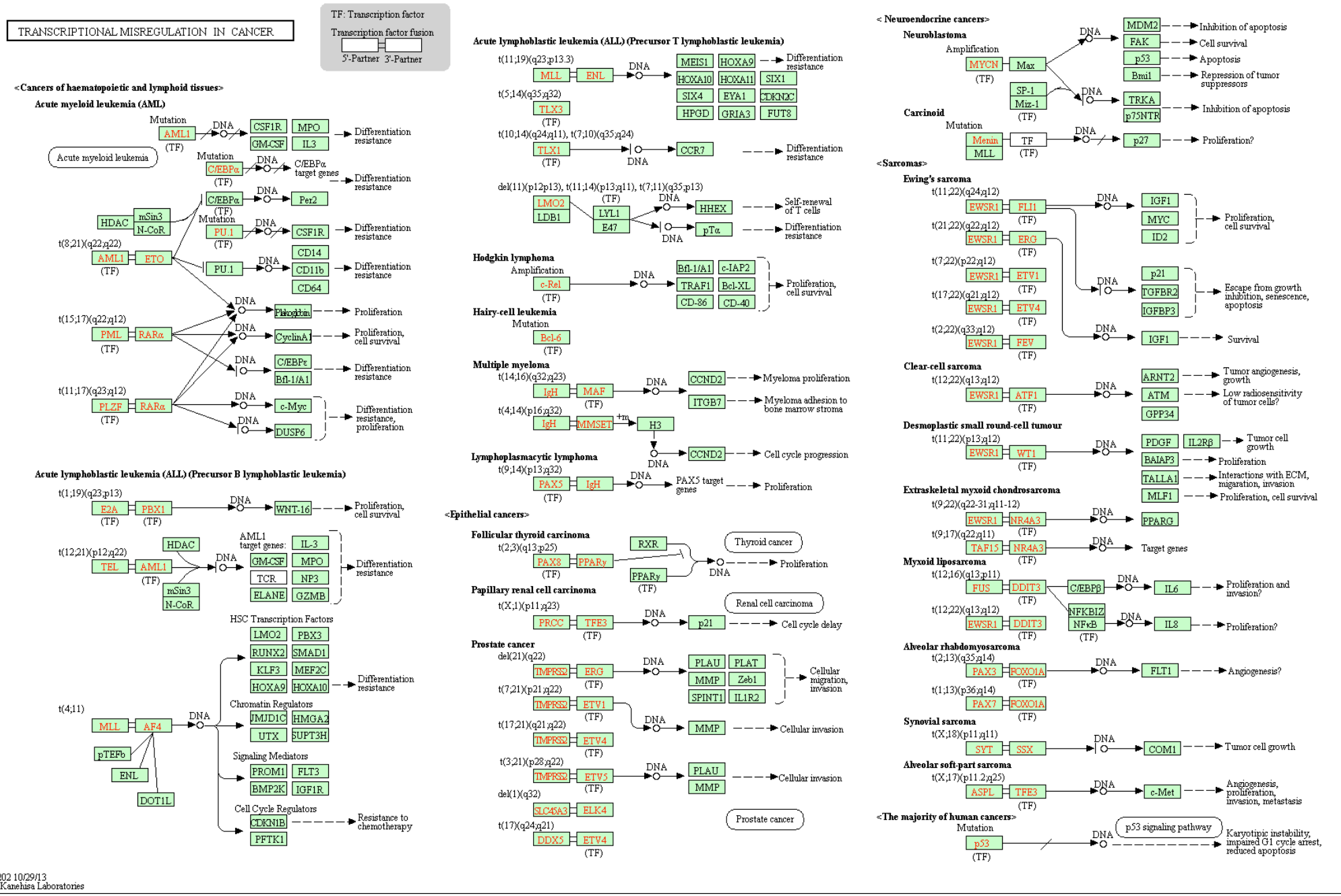
- Analytics

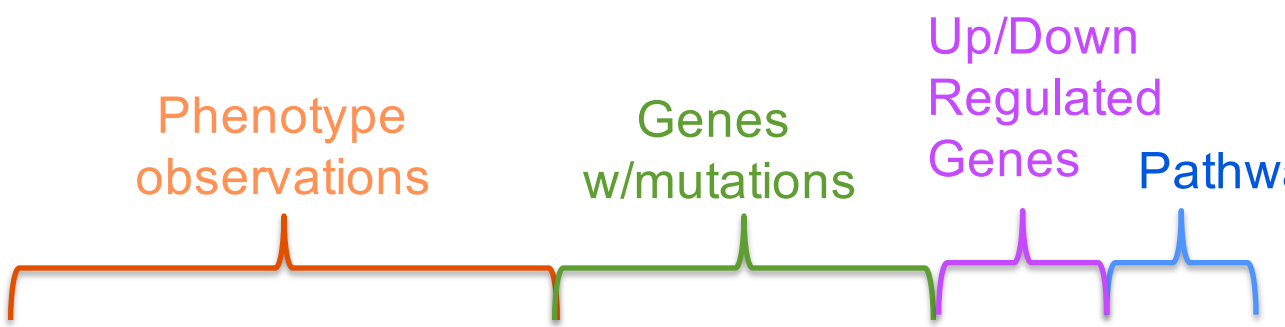
Problem 7: Cohort Identification (N+1)



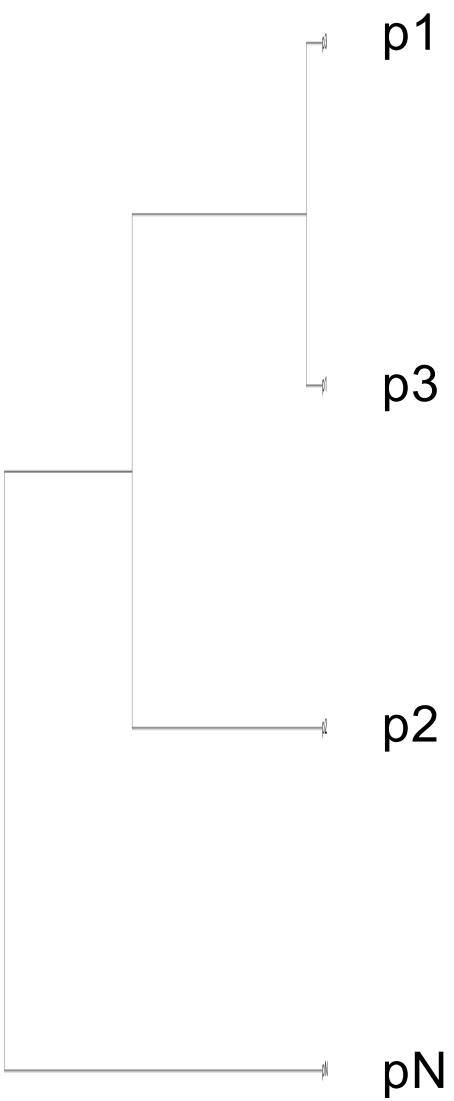
Metadata Query Builder

```
select id from patient where (
    gender = 'f',
    smoking = 'false',
    Contains(diagnosis, 'DIPG'),
    age < 10
    …
)
```



TRANSCRIPTIONAL MISREGULATION IN CANCER
TF: Transcription factor
Transcription factor fusion
5'-Partner 3'-Partner
<Cancers of haematopoietic and lymphoid tissues>
Acute myeloid leukemia (AML)
Acute myeloid leukemia
Mutation
AML1
(TF)
DNA
CSF1R
MPO
GM-CSF
IL3
Differentiation resistance
Mutation
C/EBPα
(TF)
DNA
C/EBPα target genes
Differentiation resistance
C/EBPα
(TF)
DNA
Per2
HDAC
mSin3
N-CoR
Mutation
PU.1
(TF)
DNA
CSF1R
Differentiation resistance
t(8;21)(q22;q22)
AML1
(TF)
ETO
PU.1
DNA
CD14
CD11b
CD64
Differentiation resistance
DNA
Plakoglobin
Proliferation
t(15;17)(q22;q12)
PML
(TF)
RARα
DNA
CyclinA1
Proliferation, cell survival
DNA
C/EBPε
Bfl-1/A1
Differentiation resistance
t(11;17)(q23;q12)
PLZF
(TF)
RARα
DNA
c-Myc
DUSP6
Differentiation resistance, proliferation
Acute lymphoblastic leukemia (ALL) (Precursor B lymphoblastic leukemia)
t(1;19)(q23;p13)
E2A
(TF)
PBX1
(TF)
DNA
WNT-16
Proliferation, cell survival
t(12;21)(p12;q22)
TEL
AML1
(TF)
HDAC
mSin3
N-CoR
DNA
AML1 target genes:
GM-CSF
TCR
ELANE
IL-3
MPO
NP3
GZMB
Differentiation resistance
t(4;11)
MLL
AF4
(TF)
DNA
pTEFb
ENL
DOT1L
HSC Transcription Factors
LMO2
PBX3
RUNX2
SMAD1
KLF3
MEF2C
HOXA9
HOXA10
Differentiation resistance
Chromatin Regulators
JMJD1C
HMGA2
UTX
SUPT3H
Signaling Mediators
PROM1
FLT3
BMP2K
IGF1R
Cell Cycle Regulators
CDKN1B
PFTK1
Resistance to chemotherapy
Acute lymphoblastic leukemia (ALL) (Precursor T lymphoblastic leukemia)
t(11;19)(q23;p13.3)
MLL
(TF)
ENL
DNA
MEIS1
HOXA9
HOXA10
HOXA11
SIX1
SIX4
EYA1
CDKN2C
HPGD
GRIA3
FUT8
Differentiation resistance
t(5;14)(q35;q32)
TLX3
(TF)
t(10;14)(q24;q11), t(7;10)(q35;q24)
TLX1
(TF)
DNA
CCR7
Differentiation resistance
del(11)(p12p13), t(11;14)(p13;q11), t(7;11)(q35;p13)
LMO2
LDB1
(TF)
LYL1
E47
DNA
HHEX
Self-renewal of T cells
DNA
pTα
Differentiation resistance
Hodgkin lymphoma
Amplification
c-Rel
(TF)
DNA
Bfl-1/A1
c-IAP2
TRAF1
Bcl-XL
CD-86
CD-40
Proliferation, cell survival
Hairy-cell leukemia
Mutation
Bcl-6
(TF)
Multiple myeloma
t(14;16)(q32;q23)
IgH
(TF)
MAF
DNA
CCND2
Myeloma proliferation
ITGB7
Myeloma adhesion to bone marrow stroma
t(4;14)(p16;q32)
IgH
MMSET
+m
H3
DNA
CCND2
Cell cycle progression
Lymphoplasmacytic lymphoma
t(9;14)(p13;q32)
PAX5
(TF)
IgH
DNA
PAX5 target genes
Proliferation
<Epithelial cancers>
Follicular thyroid carcinoma
t(2;3)(q13;p25)
PAX8
(TF)
PPARγ
RXR
PPARγ
(TF)
Thyroid cancer
DNA
Proliferation
Papillary renal cell carcinoma
t(X;1)(p11;q23)
PRCC
TFE3
(TF)
DNA
Renal cell carcinoma
p21
Cell cycle delay
Prostate cancer
del(21)(q22)
TMPRSS2
ERG
(TF)
DNA
PLAU
PLAT
MMP
Zeb1
SPINT1
IL1R2
Cellular migration, invasion
t(7;21)(p21;q22)
TMPRSS2
ETV1
(TF)
DNA
MMP
Cellular invasion
t(17;21)(q21;q22)
TMPRSS2
ETV4
(TF)
t(3;21)(p28;q22)
TMPRSS2
ETV5
(TF)
DNA
PLAU
MMP
Cellular invasion
del(1)(q32)
SLC45A3
ELK4
t(17)(q24;q21)
DDX5
ETV4
(TF)
Prostate cancer
< Neuroendocrine cancers>
Neuroblastoma
Amplification
MYCN
(TF)
Max
DNA
MDM2
Inhibition of apoptosis
FAK
Cell survival
p53
Apoptosis
Bmi1
Repression of tumor suppressors
SP-1
Miz-1
(TF)
DNA
TRKA
p75NTR
Inhibition of apoptosis
Carcinoid
Mutation
Menin
MLL
TF
(TF)
DNA
p27
Proliferation?
<Sarcomas>
Ewing's sarcoma
t(11;22)(q24;q12)
EWSR1
FLI1
(TF)
DNA
IGF1
MYC
ID2
Proliferation, cell survival
t(21;22)(q22;q12)
EWSR1
ERG
(TF)
t(7;22)(p22;q12)
EWSR1
ETV1
(TF)
DNA
p21
TGFBR2
IGFBP3
Escape from growth inhibition, senescence, apoptosis
t(17;22)(q21;q12)
EWSR1
ETV4
(TF)
t(2;22)(q33;q12)
EWSR1
FEV
(TF)
DNA
IGF1
Survival
Clear-cell sarcoma
t(12;22)(q13;q12)
EWSR1
ATF1
(TF)
DNA
ARNT2
Tumor angiogenesis, growth
ATM
Low radiosensitivity of tumor cells?
GPP34
Desmoplastic small round-cell tumour
t(11;22)(p13;q12)
EWSR1
WT1
(TF)
DNA
PDGF
IL2Rβ
Tumor cell growth
BAIAP3
Proliferation
TALLA1
Interactions with ECM, migaration, invasion
MLF1
Proliferation, cell survival
Extraskeletal myxoid chondrosarcoma
t(9;22)(q22-31;q11-12)
EWSR1
NR4A3
(TF)
DNA
PPARG
t(9;17)(q22;q11)
TAF15
NR4A3
(TF)
DNA
Target genes
Myxoid liposarcoma
t(12;16)(q13;p11)
FUS
DDIT3
(TF)
C/EBPβ
DNA
IL6
Proliferation and invasion?
t(12;22)(q13;q12)
EWSR1
DDIT3
(TF)
NFKBIZ
NFκB
(TF)
DNA
IL8
Proliferation?
Alveolar rhabdomyosarcoma
t(2;13)(q35;q14)
PAX3
FOXO1A
(TF)
DNA
FLT1
Angiogenesis?
t(1;13)(p36;q14)
PAX7
FOXO1A
(TF)
Synovial sarcoma
t(X;18)(p11;q11)
SYT
SSX
DNA
COM1
Tumor cell growth
Alveolar soft-part sarcoma
t(X;17)(p11.2;q25)
ASPL
TFE3
(TF)
DNA
c-Met
Angiogenesis, proliferation, invasion, metastasis
<The majority of human cancers>
Mutation
p53
(TF)
DNA
p53 signaling pathway
Karyotipic instability, impaired G1 cycle arrest, reduced apoptosis
05202 10/29/13
(c) Kanehisa Laboratories



	female	smoking	brain cancer	age < 10	PIC3CA	RARG	BRCA1	IGF2BP3	Transcriptional misregulation in cancer	…
p1	1	0	1	1	1	1	0	1	1	
p2	1	1	1	0	1	1	0	0	1	
p3	1	0	0	1	1	1	0	1	1	
…										
pN	0	0	0	1	0	0	0	0	0	

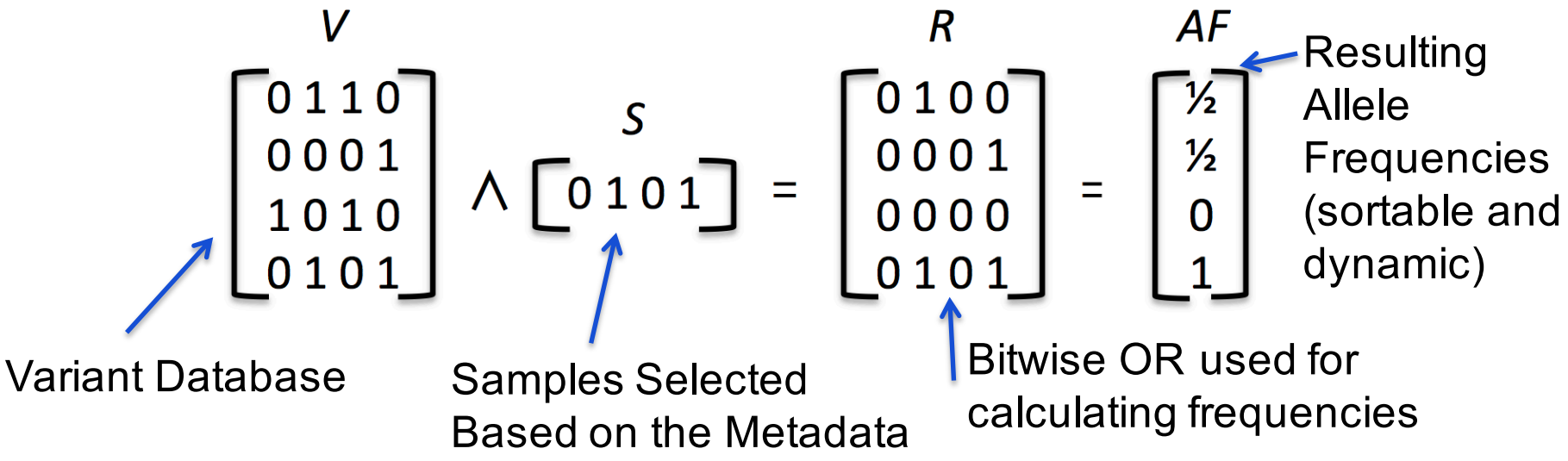


Problem 8 Dynamic Recalculation and Analytics

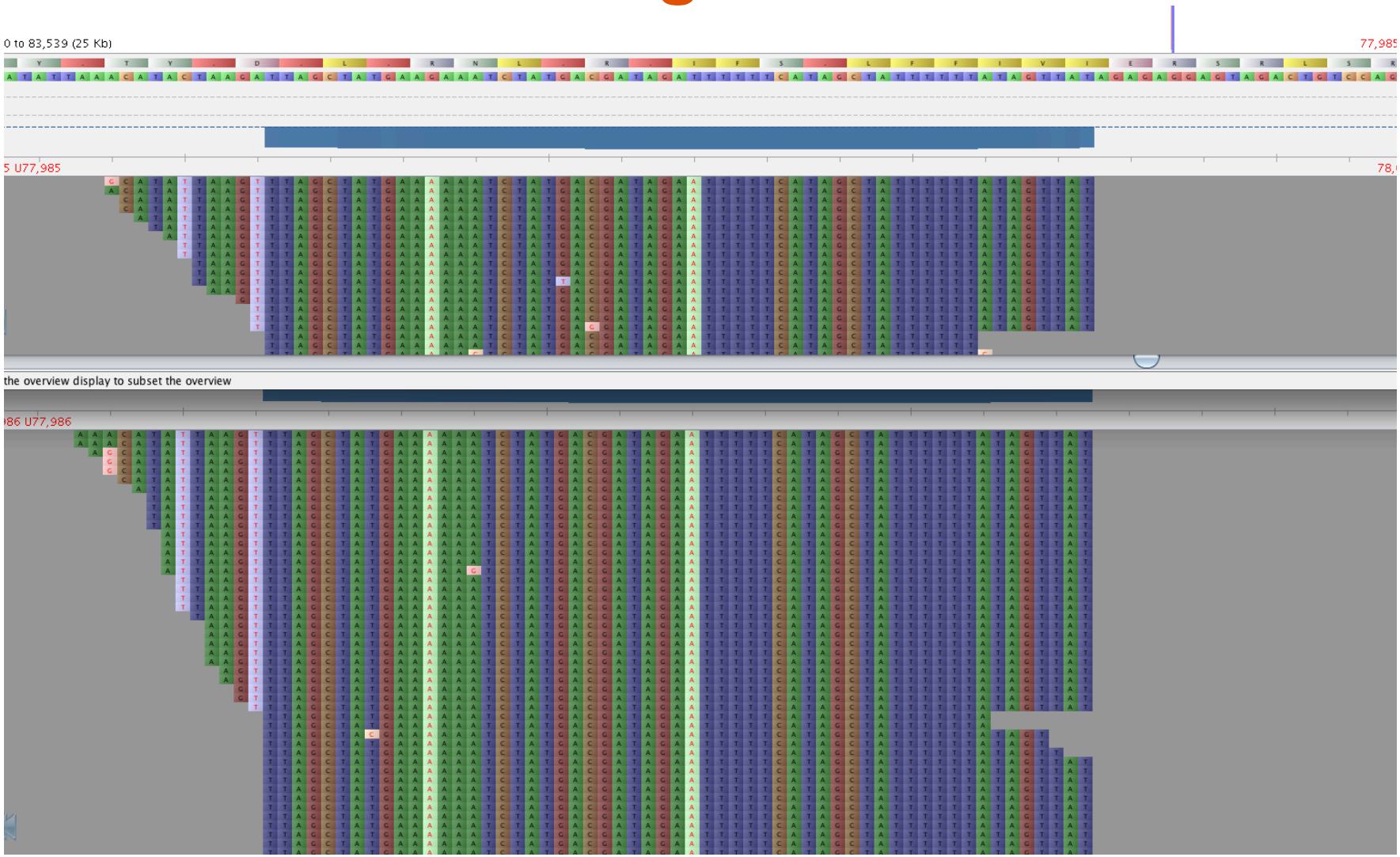
A large set of records (~70 Million) records each needs be touched in recalculating statistics.

- Statistics need to be recalculated because we often are dealing with incomplete data or incompatable technologies (e.g. gene panels versus whole genome sequencing)

After a user selects the cohort set, important statistics need to be recalculated based on the cohort set.



Different Coverage Results in Errors!



Sample 1

Sample 2

We have to normalize all of the data to make it comparable!

Each of these files come in one at a time, could be from different intuitions and could be years apart!

VCF1	VCF2	VCF3
S1	S2	S3
V1:0/1	2:1/1	V5:1/0
V2:1/1	3:0/1	
V4:1/0	4:0/1	

Variants

GVCF1	GVCF2	GVCF3
S1	S2	S3
V1	V2	V2
V2	V3	V3
V3	V4	V5
V4		

Coverage

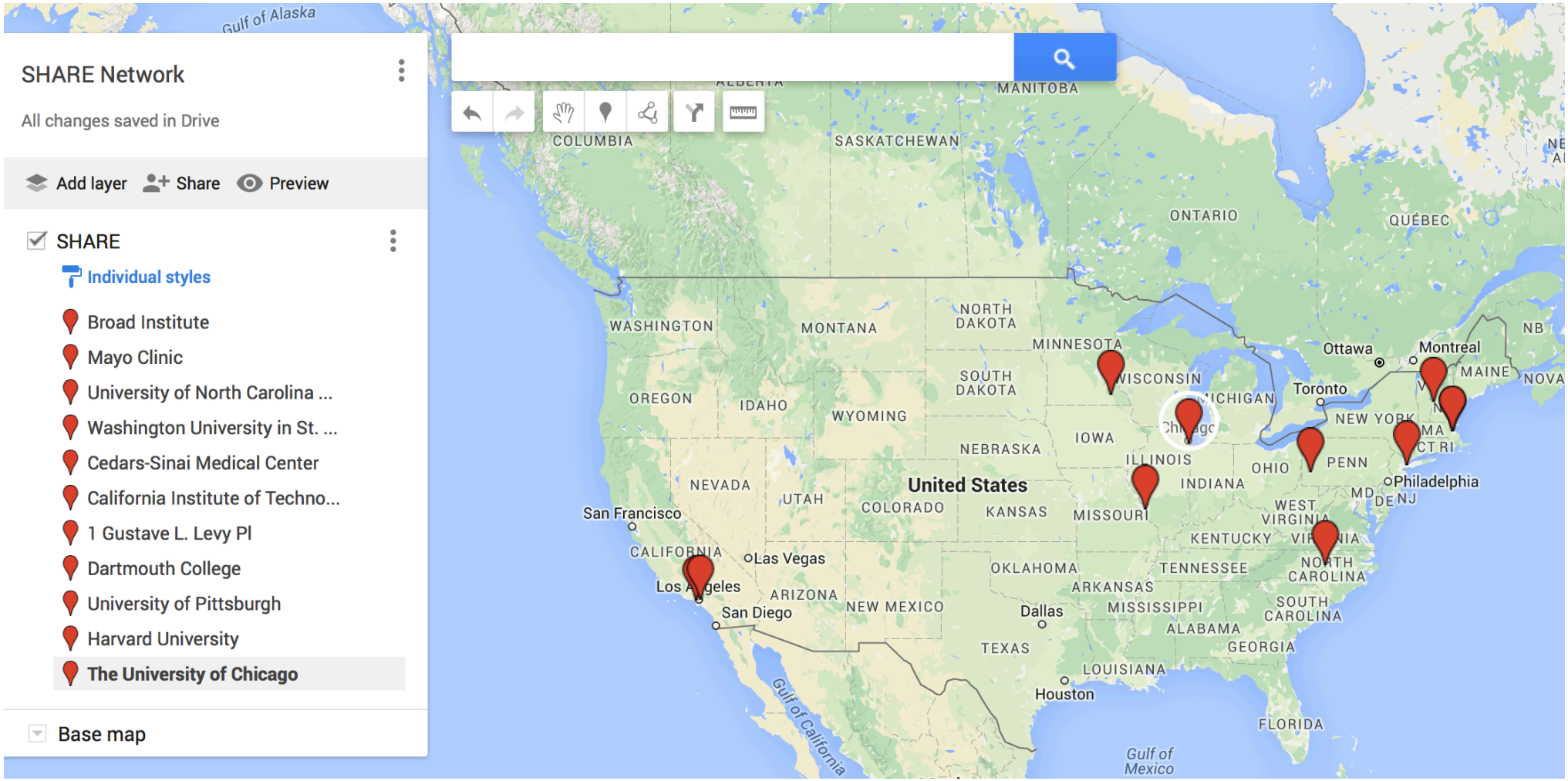
f(quality)

The Cohort Analysis requires a complete table of the following form:

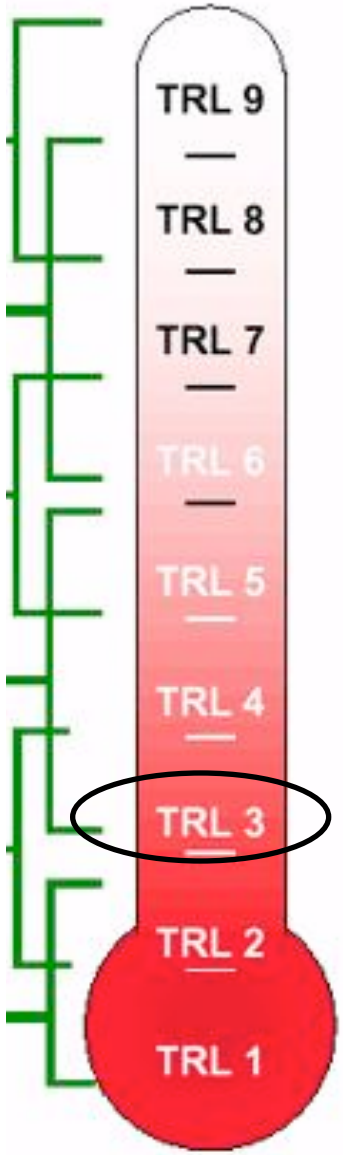
	S1	S2	S3
V1	0/1	.	.
V2	1/1	1/1	0/0
V3	0/0	0/1	0/0
V4	1/0	0/1	.
V5	.	.	1/0

Different technologies have different coverage!

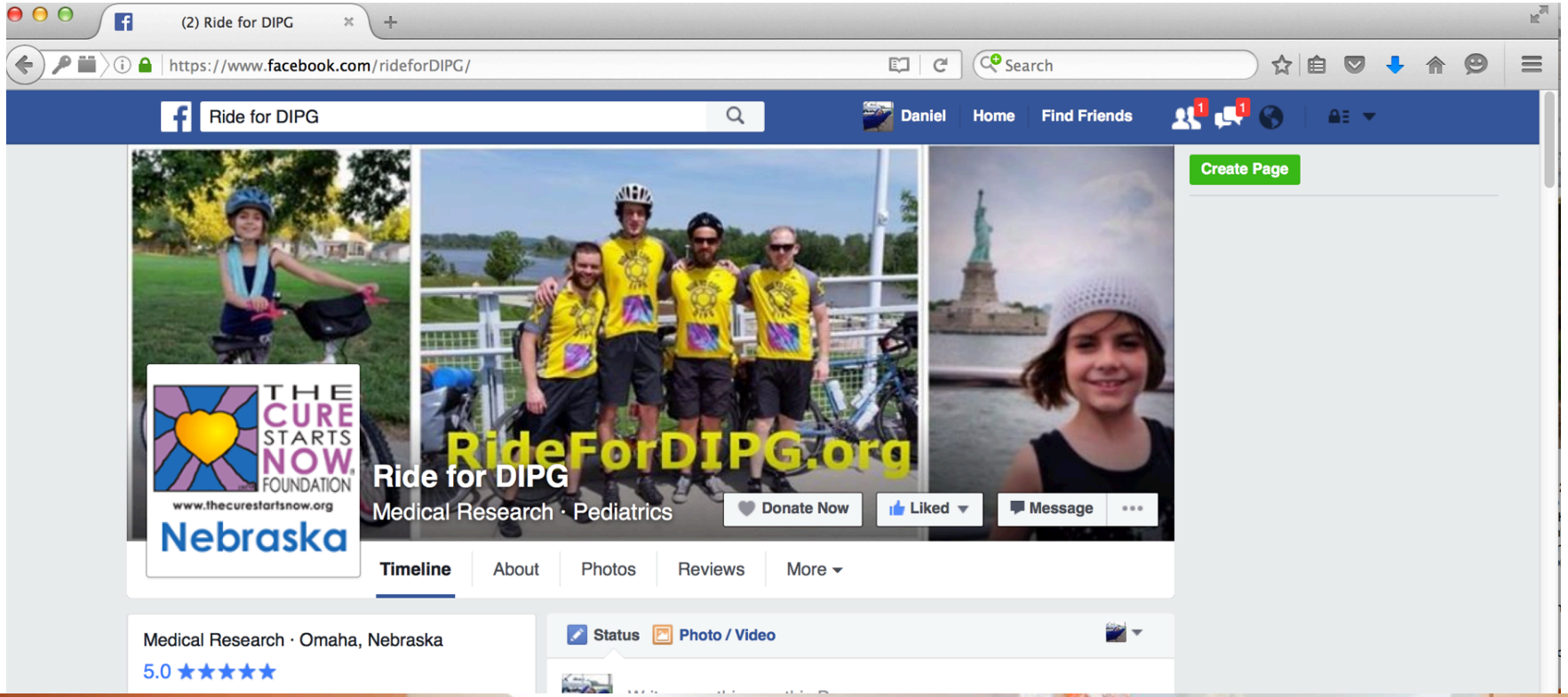
Problem 9: Data Sharing – ConsortiaDB SHARE

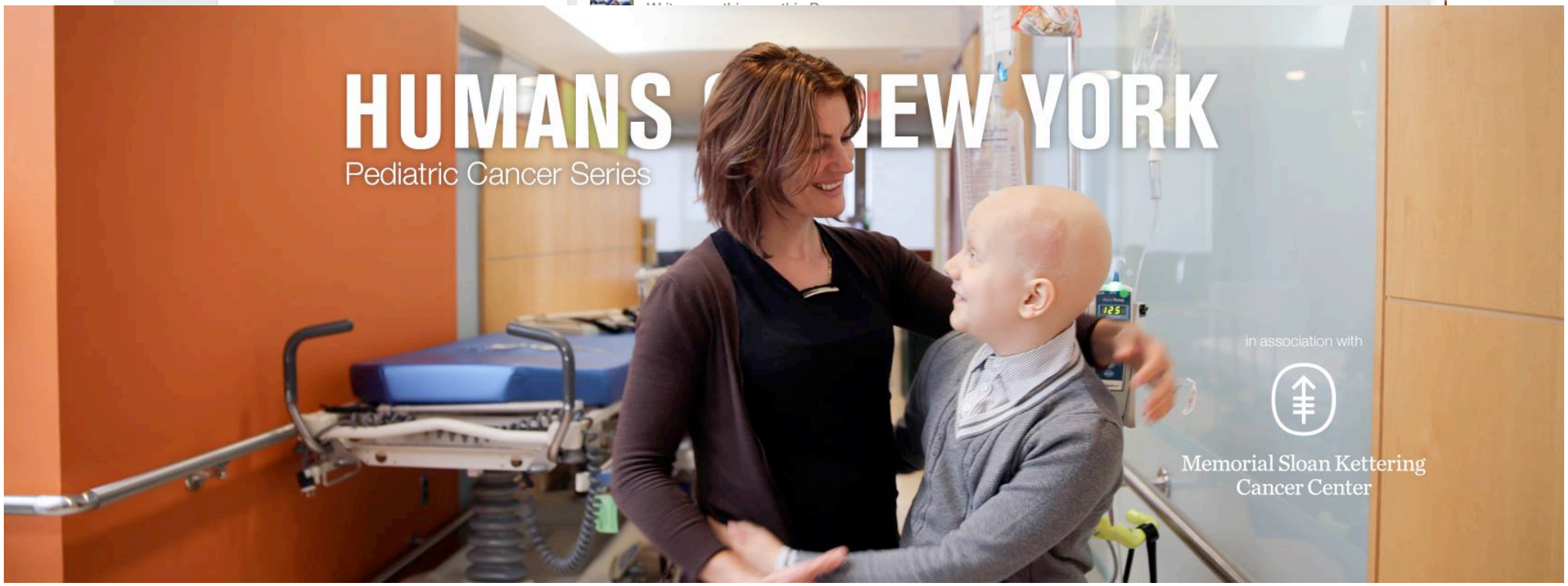


Problem	Legacy	Future State / Being Evaluated
1 – Variant Calling	GATK	GATK4 on Spark /Adam
2 - Annotation	Linux Commands - SGE	Spark
3 – Variant Filtering	MongoDB – can only filter 10,000 samples	Spark – being evaluated; could be a Spark SOLR hybrid.
4 – Clinical Oncology	Oracle	Spark / Hortonworks Stack
5 - Metadata	Elastic Search + STORM + MapReduce	Spark/SOLR Hybrid?
6 – RNA and Other Data	Linux Commands - SGE	Spark/Hadoop?
7 – Cohort Identification	DB2	Spark/Hbase Hybrid
8 – Dynamic Recalculation	Custom Distributed Java	Spark
9 – Consortium and Data Sharing	None	Spark / Hortonworks Stack



DUALIZED MEDICINE


(2) Ride for DIPG
https://www.facebook.com/rideforDIPG/
Search
Ride for DIPG
Daniel
Home
Find Friends
Create Page
THE CURE STARTS NOW FOUNDATION
www.thecurestartsnow.org
Nebraska
RideForDIPG.org
Ride for DIPG
Medical Research · Pediatrics
Donate Now
Liked
Message
Timeline
About
Photos
Reviews
More
Medical Research · Omaha, Nebraska
5.0
Status
Photo / Video


HUMANS
EW YORK
Pediatric Cancer Series
in association with
Memorial Sloan Kettering
Cancer Center

Thanks

Advanced Analytics
- Dan Blezek
- Yaxiong Lin
- Paul Bleimeyer

Natural Language Proccessing
Vinod Kaggal, Josh Pankratz, Sean Murphy, Pradip Kanjamala.

UDP
Mat Raveling, Brian Brownlow, Bob Domnick

Bioinformatics Systems
- Iain Horton
- Patrick Duffy
- Mike Meiners
- David Rider
- Matt Bockol
- Mike Kalembach
- Greg Dougherty
- David Mead

Adam Team!
Michael Heuer

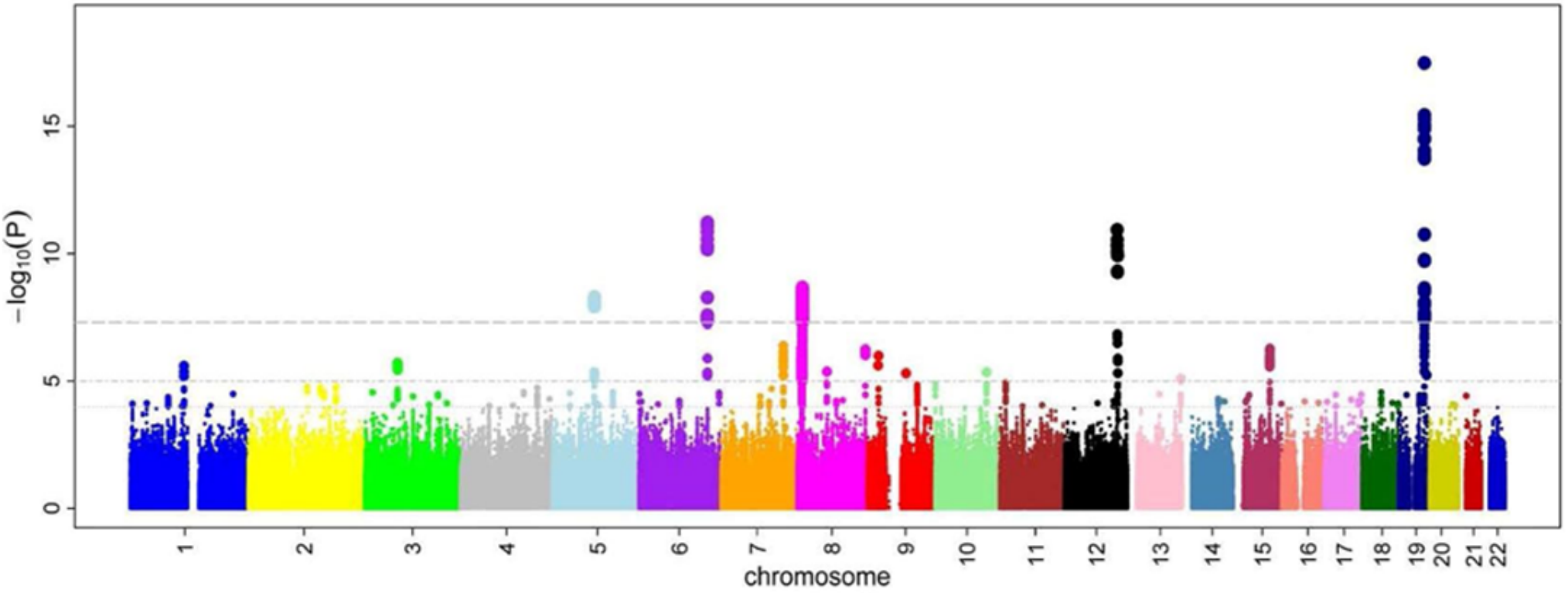
Bioinformatics Core
- JP Koche
- Steve Hart
- Raymond Moore
- Mike Zimmerman
- Dan O'Brien
- Saurabh Baheti

Advanced Analytics and Infrastructure Support
- Jason Ross

Example Personalized Medicine Case
- Jesse Shumaker
- Giselle Sholler
- Jeffry Bond

Data (!Hypothesis) Driven Discovery

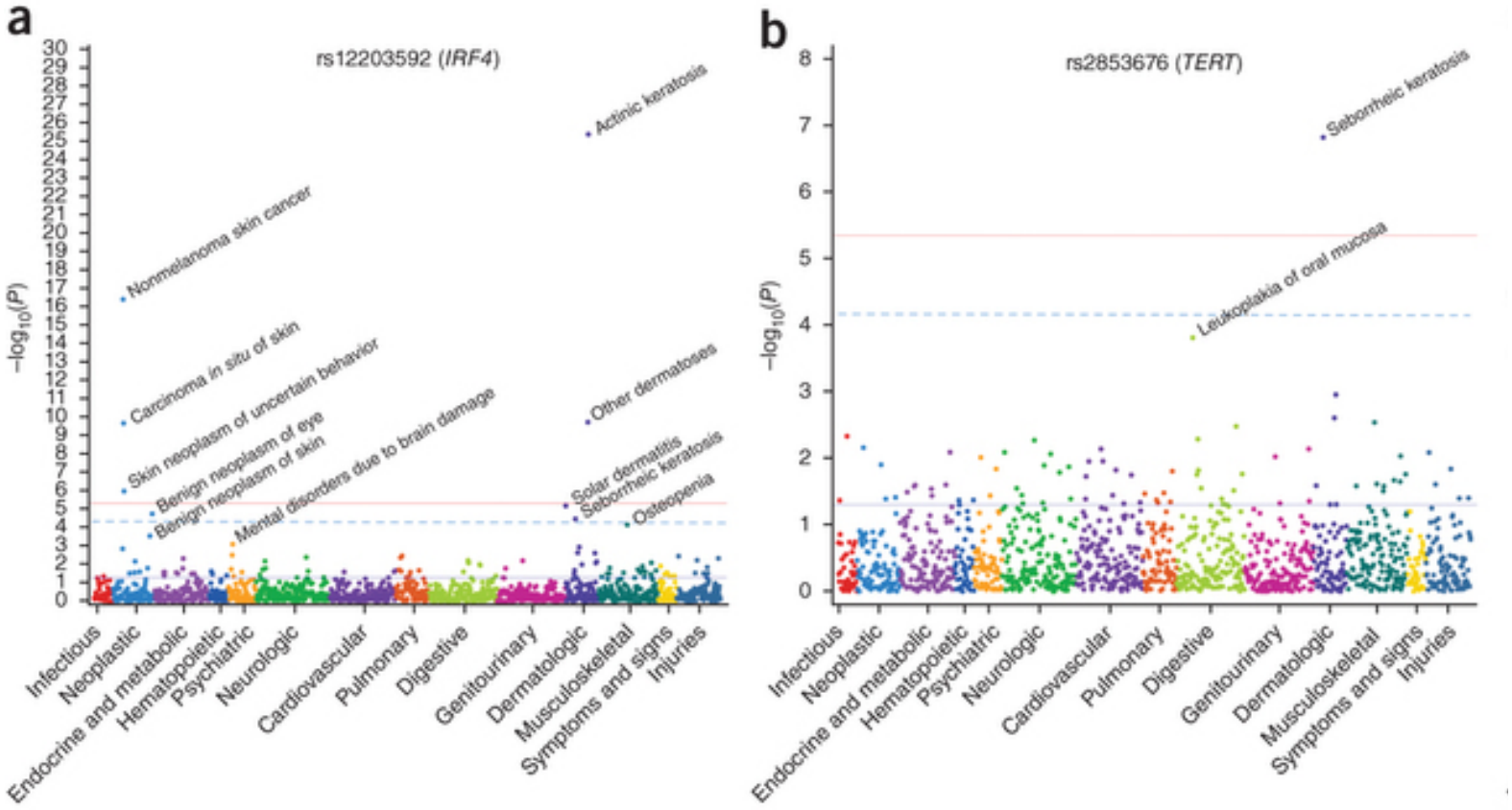
What we do today – collect data to answer a question (GWAS):



An illustration of a Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. This example is taken from a GWA study investigating microcirculation, so the tops indicates genetic variants that more often are found in individuals with constrictions in small blood vessels.[1]

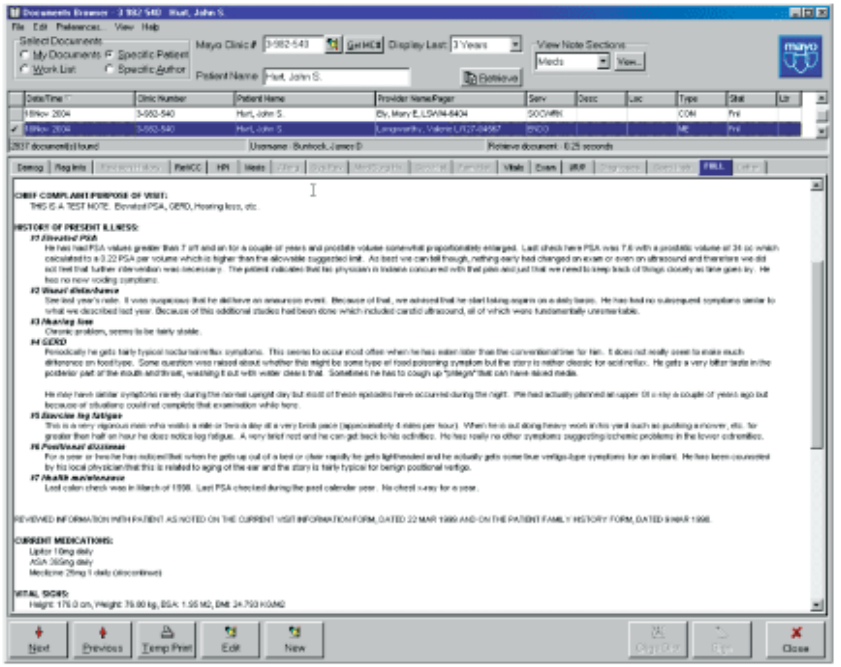
Data (!Hypothesis) Driven Discovery

What we want to do – have the data to answer a multitude of questions (PheWAS):



An illustration of a Manhattan plot depicting several strongly associated risk variants for a variety of diseases. Each dot represents a SNP, with the X-axis showing conditions and Y-axis showing association level.

Figure 2. Example of a Clinical Note



This is an ambulatory care note for a nonexistent test patient. For this study, all available inpatient and outpatient notes were used.