The Human Variant Database

Mya Warren

Michael Smith Genome Sciences Centre

Vancouver BC

Bioinformatics is Big Data

- Human genome has
  - 3 billion nucleotide bases
  - 60 thousand genes
  - 10-20 thousand proteins
- Bioinformatics takes advantage of
  - High performance computing
  - Sophisticated algorithms
  - Math/Statistics
  - Machine learning

Our mission

- Two parallel goals:
  - Personalized Oncogenomics Program

    Use patient genomics to diagnose and identify therapies for each patient's unique disease
  - Cancer research

    Find new patterns in the genomics data to identify novel targets for therapy, learn fundamental truths about cancer

Our mission

- Two parallel goals:
  - Personalized Oncogenomics Program

    Use patient genomics to diagnose and identify therapies for each patient's unique disease
  - Cancer research

    Find new patterns in the genomics data to identify novel targets for therapy, learn fundamental truths about cancer
- The database supports these goals through:
  - Fast querying and exploration of patient genomics, clinical covariates
  - Data mining and analysis of patient cohorts

HAWQ (HAdoop With Queries)

A massively parallel processing (MPP) SQL engine in Hadoop

HAWQ (HAdoop With Queries)

A massively parallel processing (MPP) SQL engine in Hadoop

- Interface with the data using PostgreSQL

HAWQ (HAdoop With Queries)

A massively parallel processing (MPP) SQL engine in Hadoop

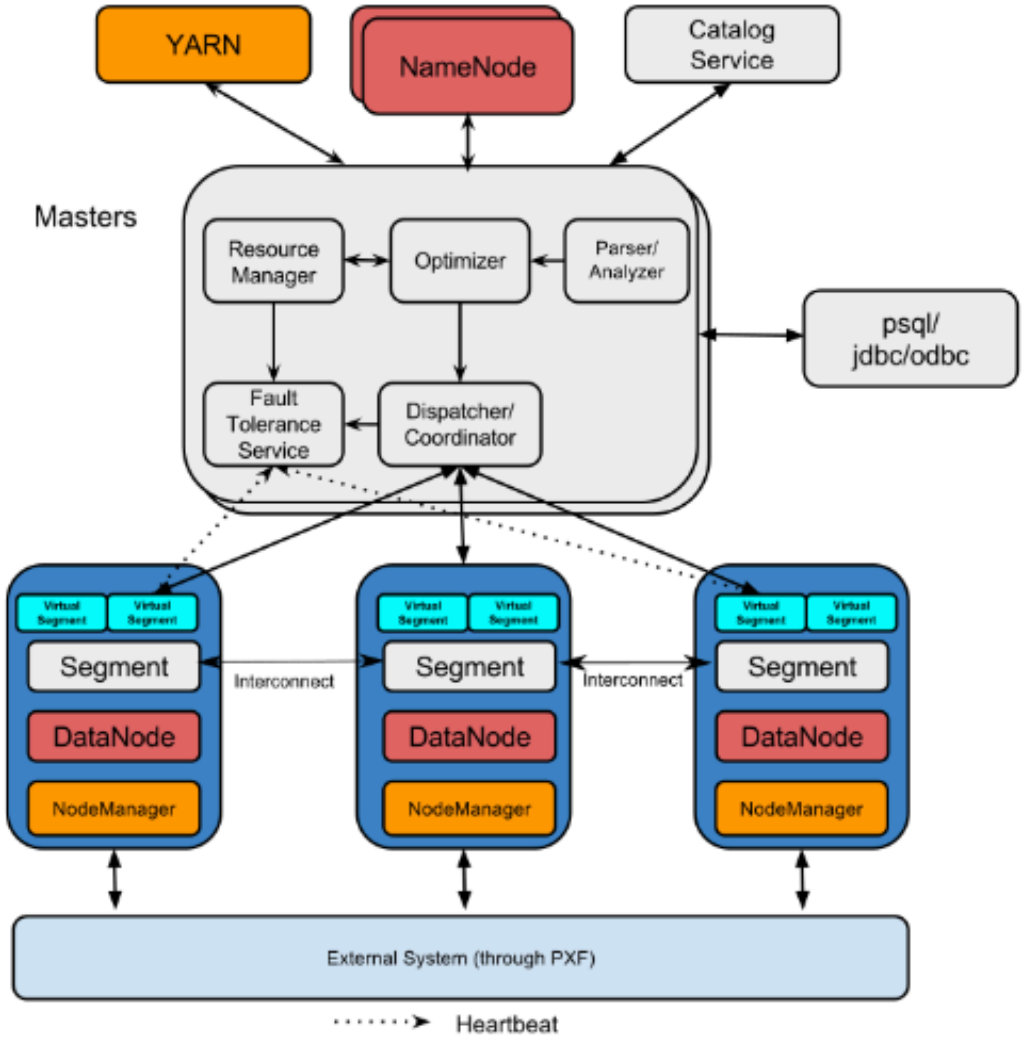
- Interface with the data using PostgreSQL
- Parallel, fault tolerant architecture for storing and processing big data

Our system

- 13 slave nodes
- 32 thread CPUs
- Total memory: 1.5 TB
- Total storage: 250 TB
- Current disk usage: 1.5 TB
- Largest table: ~10 billion rows

HAWQ Architecture

- Hadoop distributed file system (HDFS)
  - Data is chunked, replicated, distributed
- Data locality
  - Move the computation to the data
  - Data is not shared
  - HAWQ is very fast, linear scalability
- Can interface with the rest of the Hadoop ecosystem



HAWQ vs. Relational Databases

- Append-only tables
- No primary keys
- No foreign keys
- Joins are more expensive
- Extract-transform-load (ETL) optimized for large data files
  - Import raw data
  - Transform data in database

The Data

- Internally generated data + public cancer datasets (TCGA)
- 11,519 patients
- 21,591 libraries
- 31,067 analyses
- > 10 billion rows

Variants

- Raw data for
  - Unpaired/somatic SNVs and Indels
  - Germline/somatic CNVs
  - Somatic loss of heterozygosity
  - Gene expression
  - Homozygous deletions
- Post-Processed and filtered variant data

Metadata

- Library construction and sequencing
- Analysis pipeline
- Patient data
  - Demographics
  - Biopsy diagnoses
  - Drug treatment
  - Radiation treatment

Annotations

- dbSNP
- COSMIC
- ClinVar
- SnpEff
- Gene models

Coming soon

- Other internal projects
- More external data sets!
- Structural variants, miRNA...
- Disease/Drug ontologies
- Knowledgebase
- More data = better analysis!

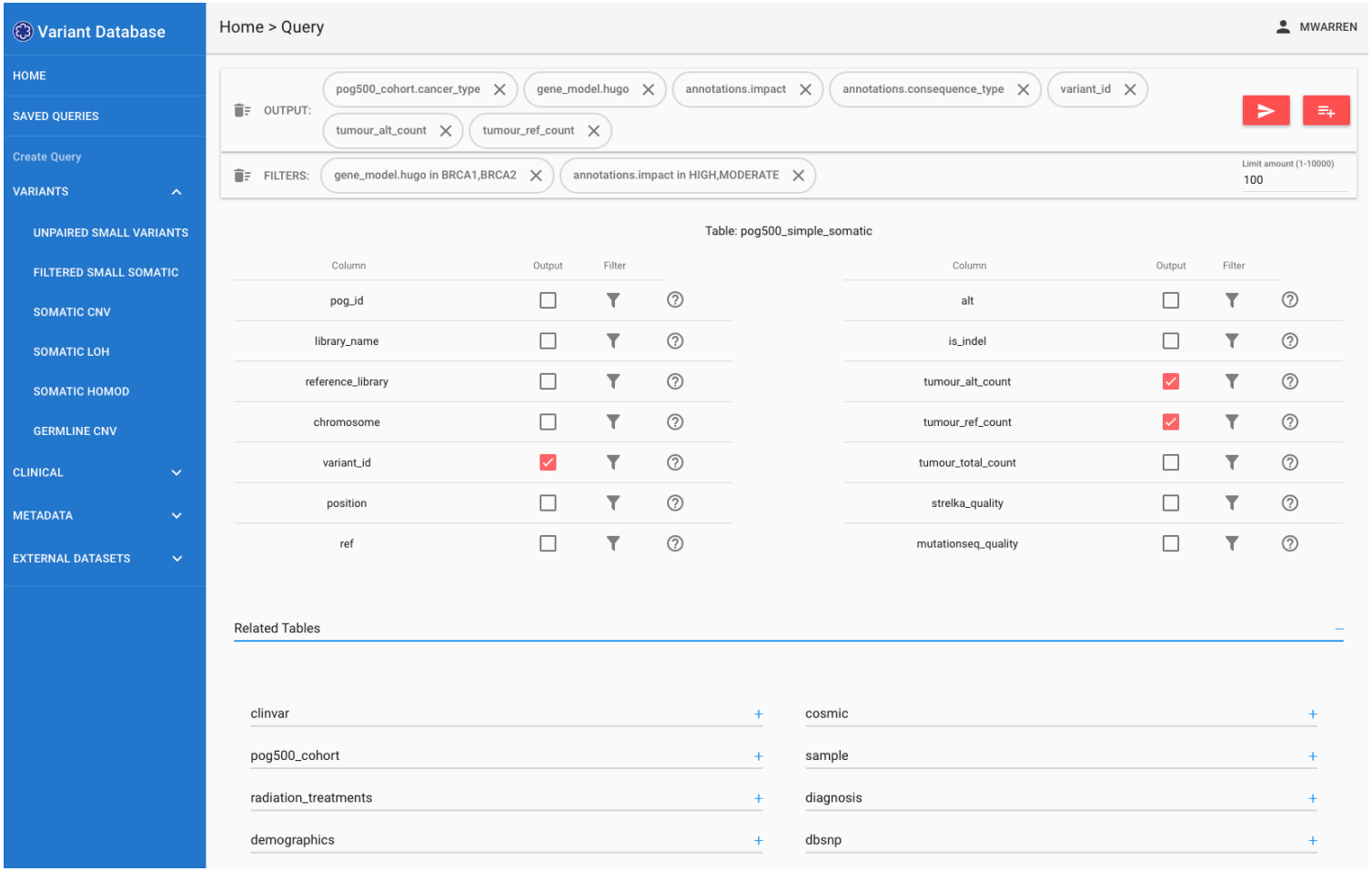
Accessing the data

- Custom queries and pipelines

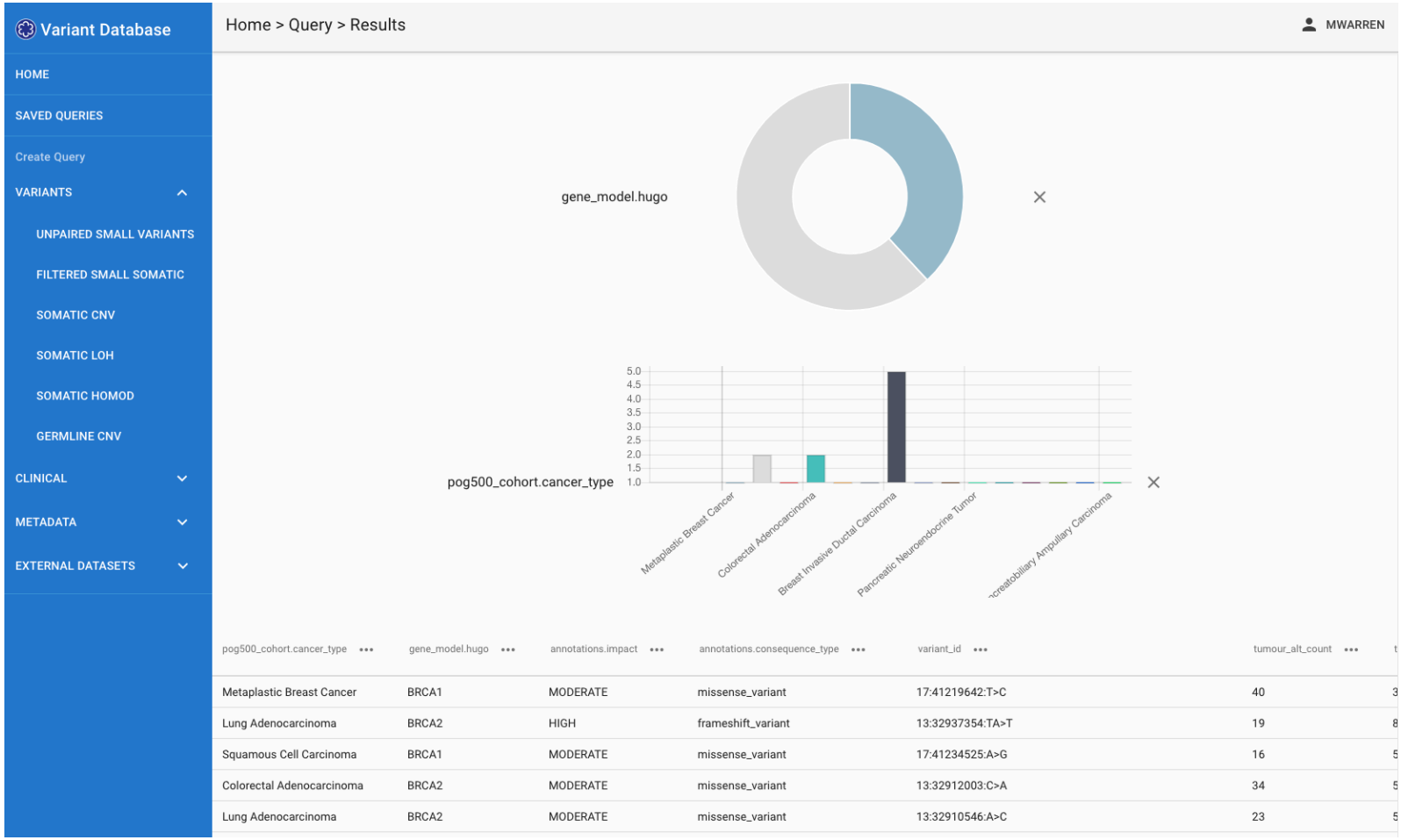
Accessing the data

- Custom queries and pipelines
- General purpose REST APIs
  - Python
  - SQL Alchemy Object Relational Model
  - Pyramid REST framework
- Web interface
  - Query
  - Filter
  - Analyze

Query selector



Results


Variant Database
HOME
SAVED QUERIES
Create Query
VARIANTS
UNPAIRED SMALL VARIANTS
FILTERED SMALL SOMATIC
SOMATIC CNV
SOMATIC LOH
SOMATIC HOMOD
GERMLINE CNV
CLINICAL
METADATA
EXTERNAL DATASETS
Home > Query > Results
MWARREN
gene_model.hugo
pog500_cohort.cancer_type
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Metaplastic Breast Cancer
Colorectal Adenocarcinoma
Breast Invasive Ductal Carcinoma
Pancreatic Neuroendocrine Tumor
Pancreatobiliary Ampullary Carcinoma
pog500_cohort.cancer_type
gene_model.hugo
annotations.impact
annotations.consequence_type
variant_id
tumour_alt_count
Metaplastic Breast Cancer BRCA1 MODERATE missense_variant 17:41219642:T>C 40
Lung Adenocarcinoma BRCA2 HIGH frameshift_variant 13:32937354:TA>T 19
Squamous Cell Carcinoma BRCA1 MODERATE missense_variant 17:41234525:A>G 16
Colorectal Adenocarcinoma BRCA2 MODERATE missense_variant 13:32912003:C>A 34
Lung Adenocarcinoma BRCA2 MODERATE missense_variant 13:32910546:A>C 23

The Future

Let the database do the work!

The Future

Let the database do the work!

- Why give up your pipeline?
  - speed
  - flexibility

Tasks that could be done on the variant database

- Annotations
- Filtering
- Statistical analysis and analytics
- Correlations
- Machine Learning

scalable, in-database analytics

Predictive Analytics Library

SUPERVISED LEARNING

Regression Models

- Cox Proportional Hazards Regression
- Elastic Net Regularization
- Generalized Linear Models
- Logistic Regression
- Marginal Effects
- Multinomial Regression
- Ordinal Regression
- Robust Variance, Clustered Variance
- Support Vector Machines

Tree Methods

- Decision Tree
- Random Forest

Other Methods

- Conditional Random Field
- Naive Bayes

UNSUPERVISED LEARNING

- Association Rules (Apriori)
- Clustering (K-means)
- Topic Modeling (LDA)

TIME SERIES

- ARIMA

MODEL EVALUATION

- Cross Validation

OTHER MODULES

- Conjugate Gradient
- Linear Solvers
- PMML Export
- Random Sampling
- Term Frequency for Text

DATA TYPES AND TRANSFORMATIONS

- Array Operations
- Dimensionality Reduction (PCA)
- Encoding Categorical Variables
- Matrix Operations
- Matrix Factorization (SVD, Low Rank)
- Norms and Distance Functions
- Sparse Vectors

STATISTICS

Descriptive

- Cardinality Estimators
- Correlation
- Summary

Inferential

- Hypothesis Tests

Other Statistics

- Probability Functions

Thanks!

Variant DB Developers

- Marcel Bernard
- Joshua Davies
- Darryl D'Souza
- Navjashan Singh
- James Zhou
- Simon Chan

PIPE/BioApps/LIMS

- Morgan Bye
- Karen Eddy
- Patrick Plettner

Systems

- Hansen Wong
- Rudy Zhou
- Lance Bailey

Brandon Pierce

Richard Corbett

Eric Chuah

Yussanne Ma