Working Assumptions of i2b2 Data

Shawn Murphy
Vivian Gainer

Outline of Workshop #1

- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

Outline of Workshop #1

- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

Useful data for research

- Defining the customers
- Finding the data most useful to your customers
- Avoiding data scope creep
- Data challenges to be avoided
- Understanding limitations of the data

Defining the customers

- Those gathering research cohorts
- Those using data to find associations
- Those performing operations research

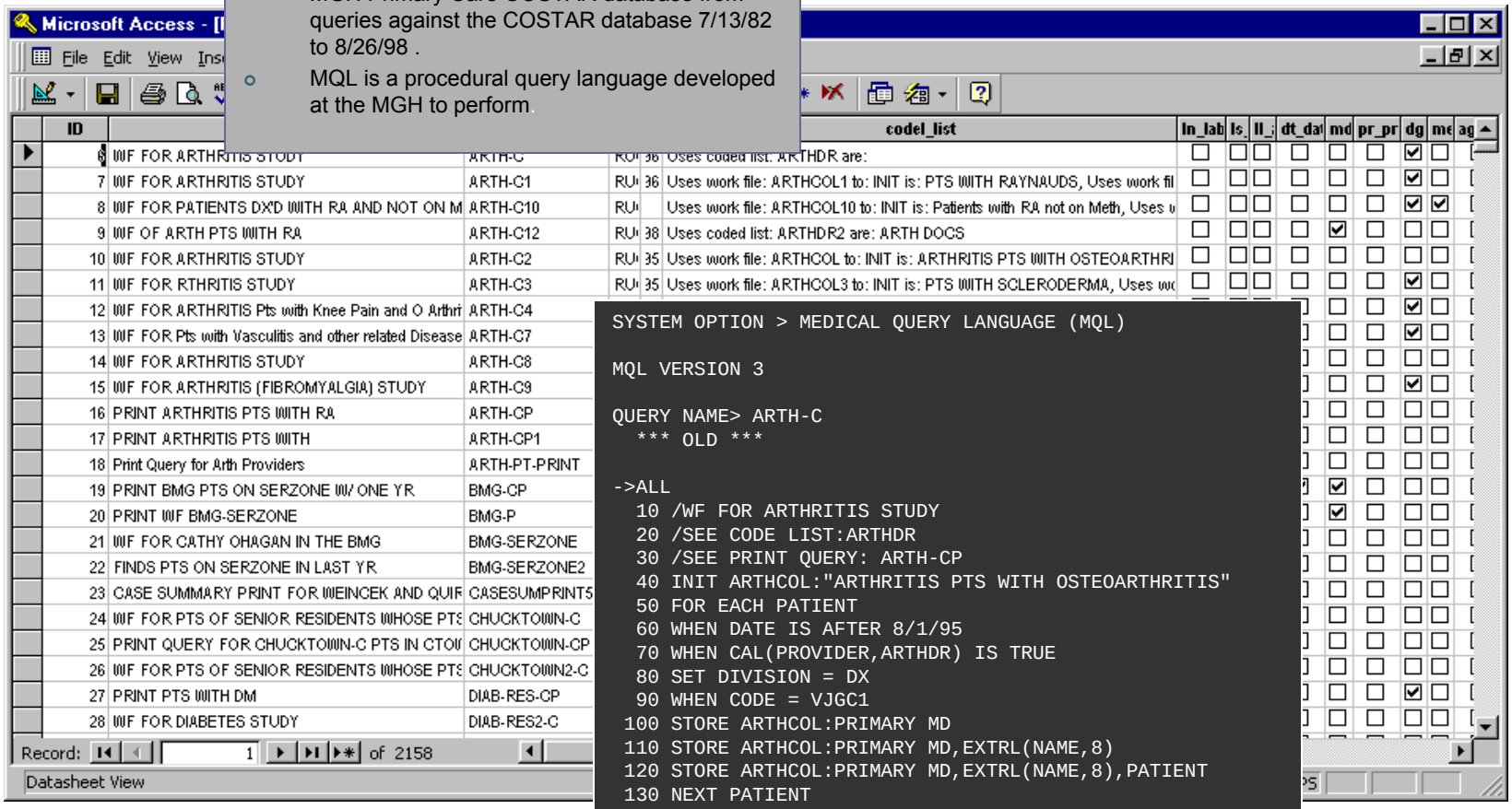
- Specificity/Sensitivity
- Need for completeness
- Accuracy of values

What data is most useful

- Asking individual clinical researchers what data would be useful to them is not necessarily the best approach, they tend to be focused on a small problem space.
- Another approach is to gather evidence from databases they have attempted to use.

Example – patient cohorts

- 642 MQL queries used to find patient cohorts for research studies were analyzed from the MGH Primary Care COSTAR database from queries against the COSTAR database 7/13/82 to 8/26/98 .
- MQL is a procedural query language developed at the MGH to perform.




```
SYSTEM OPTION > MEDICAL QUERY LANGUAGE (MQL)

MQL VERSION 3

QUERY NAME> ARTH-C
  *** OLD ***

->ALL
  10 /WF FOR ARTHRITIS STUDY
  20 /SEE CODE LIST:ARTHDR
  30 /SEE PRINT QUERY: ARTH-CP
  40 INIT ARTHCOL:"ARTHRITIS PTS WITH OSTEOARTHRITIS"
  50 FOR EACH PATIENT
  60 WHEN DATE IS AFTER 8/1/95
  70 WHEN CAL(PROVIDER,ARTHDR) IS TRUE
  80 SET DIVISION = DX
  90 WHEN CODE = VJGC1
 100 STORE ARTHCOL:PRIMARY MD
 110 STORE ARTHCOL:PRIMARY MD,EXTRL(NAME,8)
 120 STORE ARTHCOL:PRIMARY MD,EXTRL(NAME,8),PATIENT
 130 NEXT PATIENT
->
```

Example – patient cohorts

- 642 MQL queries used to find patient cohorts for research studies were analyzed from the MGH Primary Care COSTAR database from 7/13/82 to 8/26/98 .
- MQL is a procedural query language developed at the MGH to perform queries against the COSTAR database.
- Major data groups were analyzed assuming demographics, providers, and encounter detail are present.

 56	**Coded Diagnosis**
 36	**Coded Medications**
 6	**Coded Procedures**
 11	**Lab values (numeric)**
 1	**Lab values (short text)**
 2	**Lab tests (long reports)**
 7	**Clinical reports**

Encounter Dates	80%
Age	70%
Gender	60%
Practitioner	40%

Percent usage of various data classes

Avoid data scope creep

- Greatest single risk to most data gathering projects

Data Challenges

- No unified identifier
  - Even available identifiers are often in constant flux.
- No coding system used with data
  - Coding system is used, but not managed appropriately
- Values are not properly constrained

Data Challenges

- No method to obtain data updates
  - Update flag may exist in source system, but not properly managed

Understanding limitations of the data

- The accuracy of coding varies with different diagnoses, i.e., it may be excellent for some diagnosis and poor for others. Generally this is highly correlated with the representation of the diagnosis in the coding system (Valinsky et al, 1999)
  - “Any epidemiological research that uses surgical complication codes from operative admissions, particularly in the absence of a specific ICD9CM code, will lead to significantly underestimating the prevalence of complications”

Understanding limitations of the data

- Adding clinical correlates of a diagnosis will greatly increase the specificity for that diagnosis. (Singh et al 2004)
  - Diagnosis by ICD code 714 had 100% sensitivity, but specificity was only 55% because of a false-positive rate of 34%. The addition of a positive RF and/or a DMARD prescription to ICD code 714 dramatically improved specificity to 83–97% and positive predictive value to 81–97%; however, sensitivity decreased to 76–88%

Understanding limitations of the data

- It is rare to find a code in medical record systems that represents the negation of a concept. Therefore, it is not possible to distinguish between the true absence of a disease and the absence of the recording of the disease.
  - Specific provider and clinic data completeness patterns can be used to help address this problem.

Outline of Workshop #1

- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

Data Commonly Available

- Decision Support Systems
- Look for systems feeding data to Health Consortium

Data Commonly Available

- Billing Systems

Data Commonly Available

- Patient Registration Systems

Data Commonly Available

- Electronic Medical Record Systems

Data Commonly Available

- Federal and State registries
  - Death indexes

Data Commonly Available

- Claims data from insurance companies

Data Commonly Available

- Text reports

Data Commonly Available

- Images

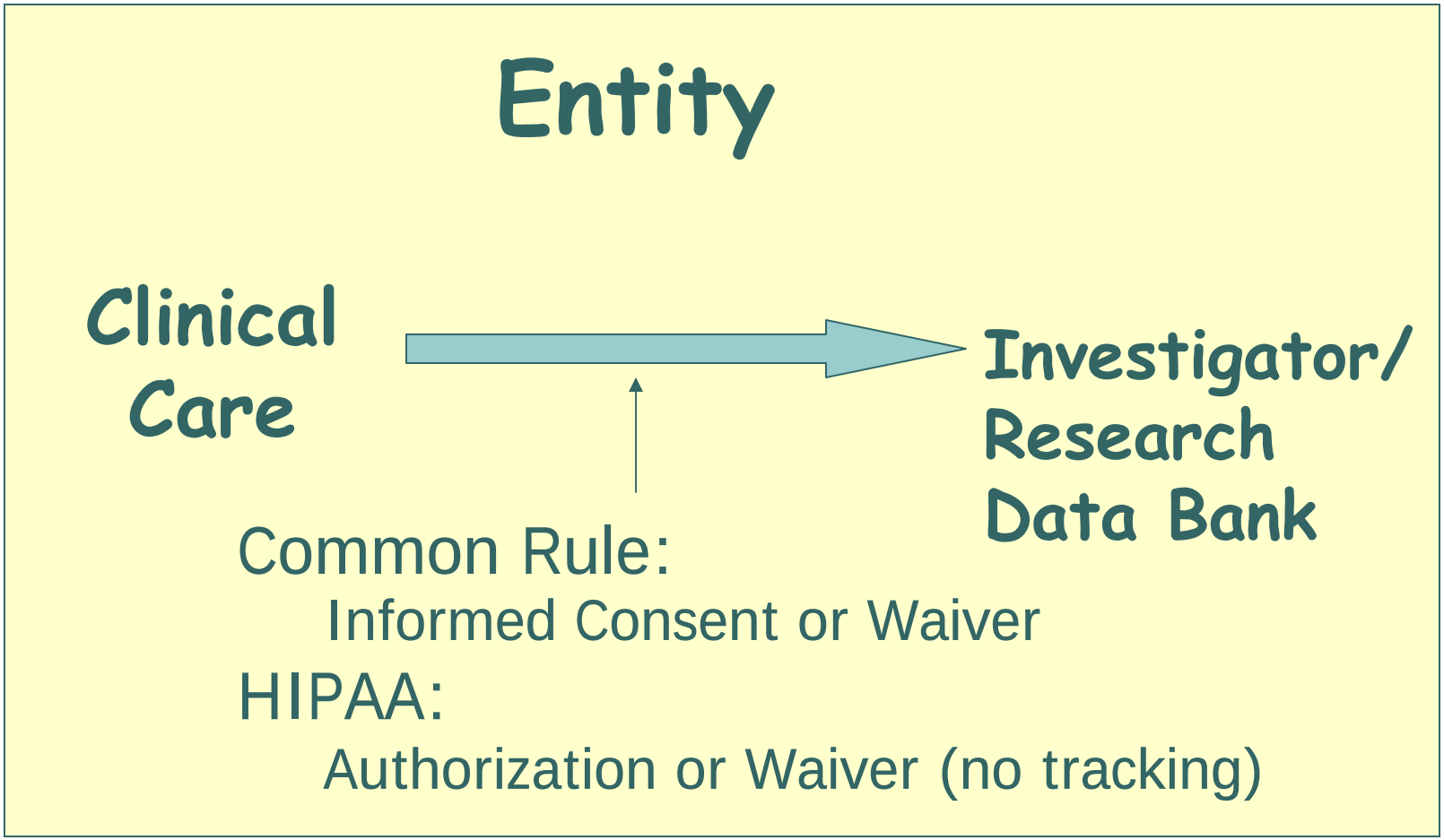
Outline of Workshop #1

- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

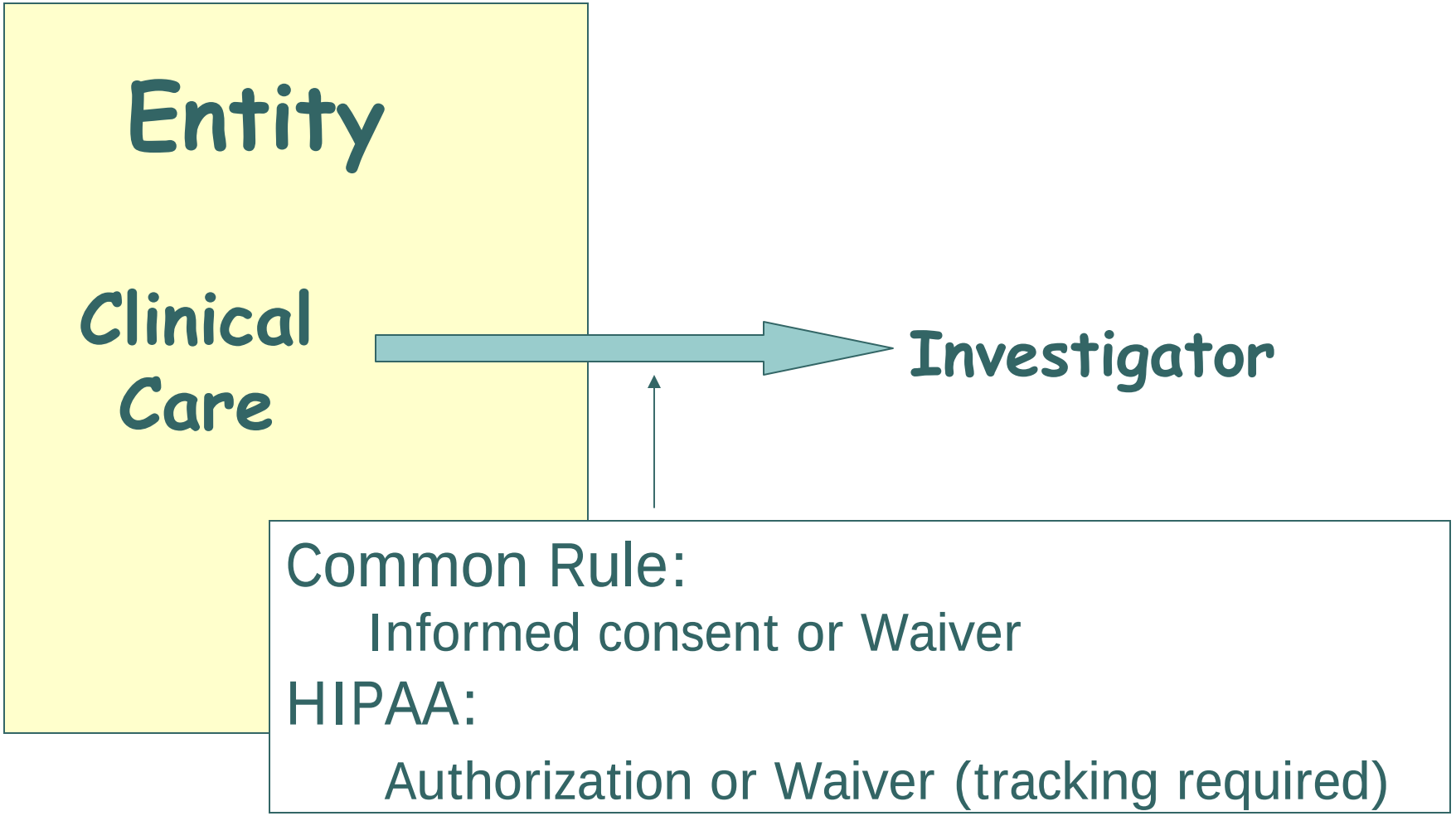
Privacy of clinical data

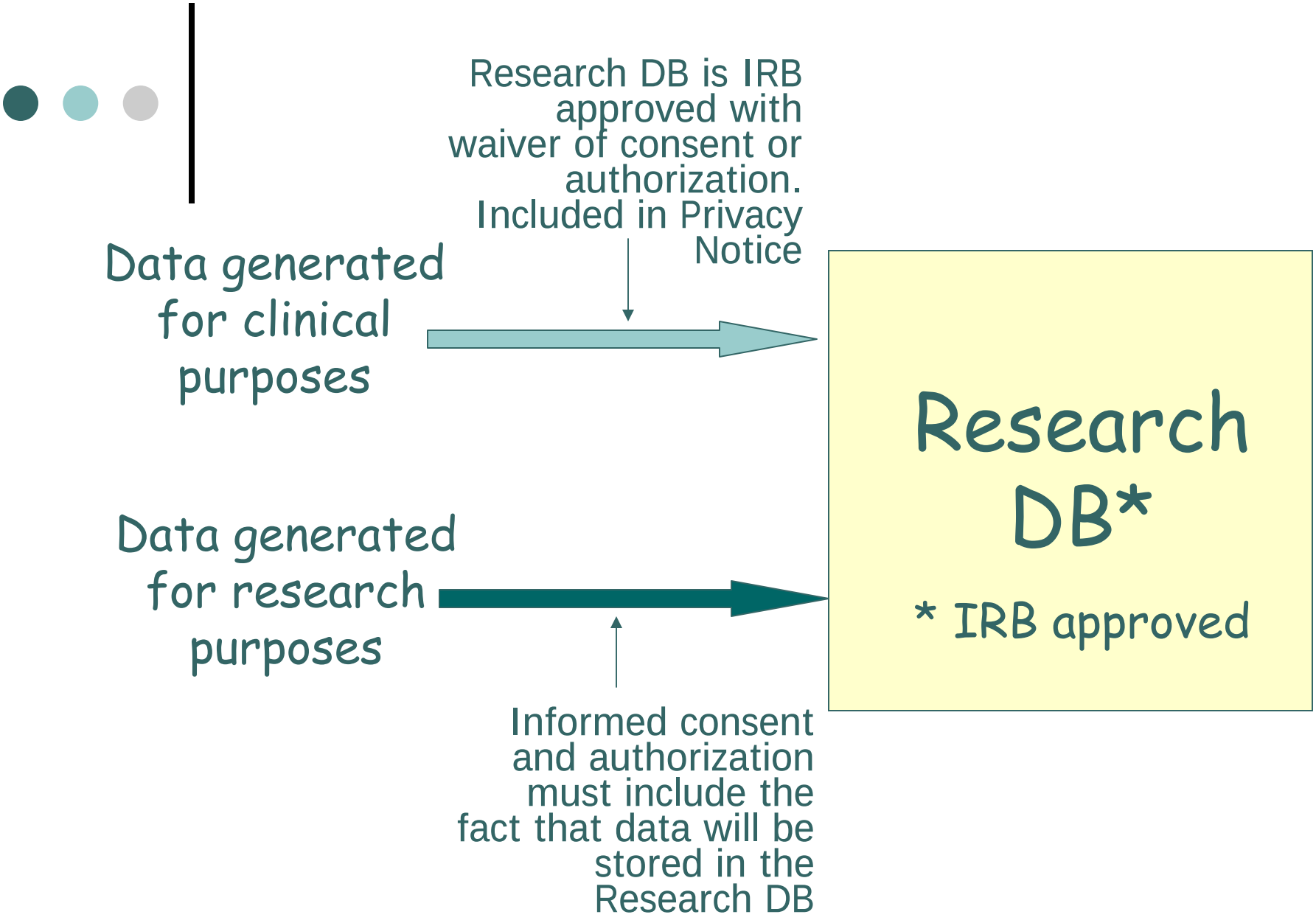
- Policy
- Methods

Flow of Identifiable Information For Research



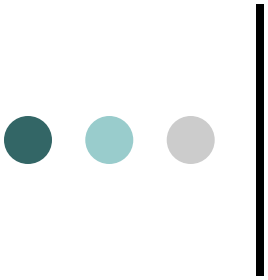
Flow of Identifiable Information For Research




Research DB is IRB approved with waiver of consent or authorization. Included in Privacy Notice
Data generated for clinical purposes
Data generated for research purposes
Informed consent and authorization must include the fact that data will be stored in the Research DB
Research DB*
* IRB approved

Privacy of clinical data

- Policy
- Methods



- Mapping tables

Outline of Workshop #1

- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

Data Model: Data Requirements

- **Integration** of data from distributed and differently structured databases in order to perform comprehensive analyses.
- **Separation** of data used for research from daily operational or transactional data.
- **Standardization** of a model across systems.
- **Ease** of use by end-users.

Dimensional Modeling

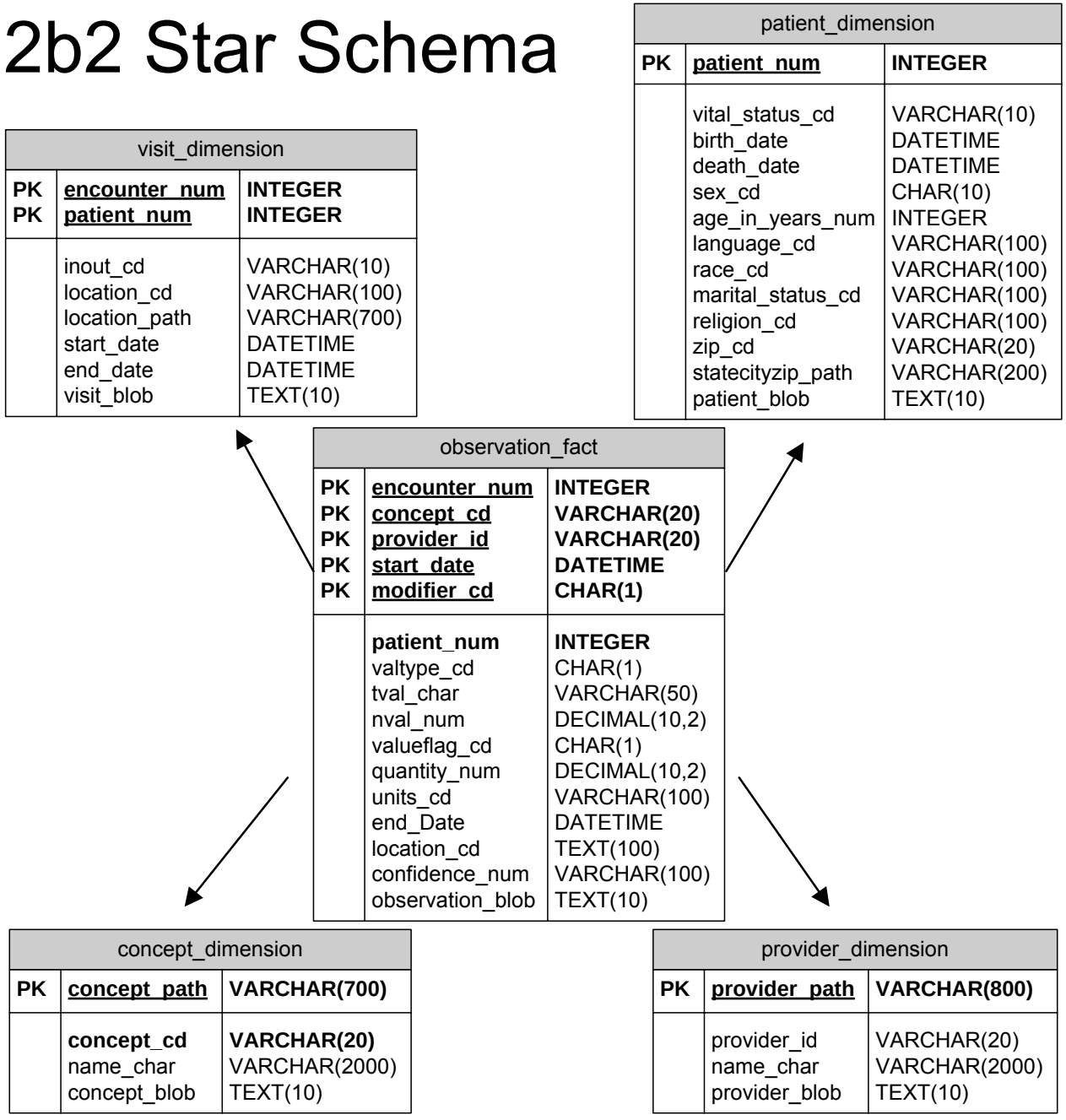
1. **FACTS** - the quantitative or factual data being queried.

2. **DIMENSIONS** – groups of hierarchies and descriptors that define the facts.

Star Schema

One fact table surrounded radially by numerous dimension tables.

i2b2 Star Schema



i2b2 Fact Table

- In i2b2, a fact is an observation on a patient.
- Examples of FACTS:
  - Diagnoses
  - Procedures
  - Health History
  - Genetic Data
  - Lab Data
  - Provider Data
  - Demographics Data
- An observation is not necessarily the same thing as an event

i2b2 Dimension Tables

- Dimension tables contain descriptive information about facts.
- In i2b2 there are four dimension tables

concept_dimension
provider_dimension
visit_dimension
patient_dimension

Indexes

- Very large data warehouses and marts require many indexes for good performance. Use as many indexes as necessary for covering virtually any query

- Consider adding a clustered index (SQL Server) to any table in a data warehouse that needs to produce sorted results.

observation_fact indexes

sp_helpindex observation_fact

index_name	index_description	index_keys
PK_Observation_Fact	nonclustered, unique, primary key located on PRIMA...	Encounter_Num, Concept_Cd, Provider_Id, Start_Date, Modifier_Cd
XIE2Observation_Fact	clustered located on PRIMARY	Concept_Cd
XIE3Observation_Fact	nonclustered located on PRIMARY	Patient_Num, Encounter_Num, Concept_Cd, Provider_Id, Start_D...
XIE4Observation_Fact	nonclustered located on PRIMARY	Start_Date, Patient_Num
XIE5Observation_Fact	nonclustered located on PRIMARY	Modifier_Cd
XIE6Observation_Fact	nonclustered located on PRIMARY	Provider_Id, Patient_Num
XIE7Observation_Fact	nonclustered located on PRIMARY	Patient_Num, Encounter_Num
XIE8Observation_Fact	nonclustered located on PRIMARY	TVal_Char, NVal_Num

*

*

Patient_Num, Encounter_Num, Concept_Cd, Provider_Id, Start_Date, Modifier_Cd, ValType_Cd, TVal_Char, NVal_Num, ValueFlag_Cd, Quantity_Num, Units_Cd, End_Date, Location_Cd, Confidence_Num

concept_dimension indexes

sp_helpindex concept_dimension

Index name	Index description	Index keys
PK_Concept_Dimension	clustered, unique, primary key located on PRIMARY	Concept_Path
XIEConcept_Dimension1	nonclustered located on PRIMARY	Concept_Cd

visit_dimension indexes

sp_helpindex visit_dimension

Index_name	Index_description	Index_keys
PK_Visit_Dimension	clustered, unique, primary key located on PRIMARY	Encounter_Num, Patient_Num
XIE3Visit_Dimension	nonclustered located on PRIMARY	Patient_Num, Encounter_Num
XIE4Visit_Dimension	nonclustered located on PRIMARY	Start_Date
XIE5Visit_Dimension	nonclustered located on PRIMARY	End_Date

patient_dimension indexes

sp_helpindex patient_dimension

Index_name	Index_description	Index_keys
PK_Patient_Dimension	clustered, unique, primary key located on PRIMARY	Patient_Num
XIE4Patient_Dimension	nonclustered located on PRIMARY	Age_In_Years_Num
XIE6Patient_Dimension	nonclustered located on PRIMARY	Race_Cd

provider_dimension indexes

sp_helpindex provider_dimension

Index_name	Index_description	Index_keys
PK_Provider_Dimension	clustered, unique, primary key located on PRIMARY	Provider_Path, Provider_Id
XIEProvider_Dimension1	nonclustered located on PRIMARY	Provider_Id

Values

Valtype_cd	either N for numeric or T for text
Tval_char	if valtype_cd = ‘T’, then the text value goes here.
	if valtype_cd = ‘N’, then tval_char can be ‘E’ for equals, G for greater than, L for less than
Nval_num	if valtype_cd = ‘N’, then the text value goes here
Valueflag_cd	Flag (for high or low values, for example)

Example: Lab Test Values

```
select o.concept_cd, name_char, valtype_cd, tval_char,
nval_num, valueflag_cd, units_cd
from observation_fact o join concept_dimension  c
on o.concept_cd = c.concept_cd
where valtype_cd = 'N‘
```

concept_...	name_char	valtype_cd	tval_char	nval_num	valueflag_cd	units_cd
BC1-20	Alt/gpt (Test:bc1-20)	N	E	6.00000	L	u/l
BC1-21	Ast/got (Test:bc1-21)	N	E	16.00000	@	u/l
BC1-24	Alk phos (Test:bc1-24)	N	E	107.00000	@	u/l
BC1-39	Albumin (Test:bc1-39)	N	E	4.20000	@	g/dl
BC1-7	Creatinine (Test:bc1-7)	N	E	0.70000	@	mg/dl
BC1-10	Chloride (Test:bc1-10)	N	E	106.00000	@	mmol/l
BC1-106	B12 (Test:bc1-106)	N	E	409.00000	@	pg/ml
BC1-11	Total co2 (Test:bc1-11)	N	E	26.00000	@	mmol/l
BC1-110	Ferritin (Test:bc1-110)	N	E	42.00000	@	ug/l
BC1-136	Vldl (Test:bc1-136)	N	E	12.00000	@	mg/dl
BC1-19	Anion gap (Test:bc1-19)	N	E	10.00000	@	mmol/l
BC1-20	Alt/gpt (Test:bc1-20)	N	E	9.00000	@	u/l

i2b2 Metadata

Meta

from Greek: μετά = "after", "beyond", "with“, "change”

Relationship of Metadata to Star Schema

- Star Schema contains one fact and many dimension tables.
- Concepts in these tables are defined in a separate metadata table or tables.
- The structure of the metadata is integral to the visualization of concepts as well as for querying the data.
- All metadata tables have the same basic structure.

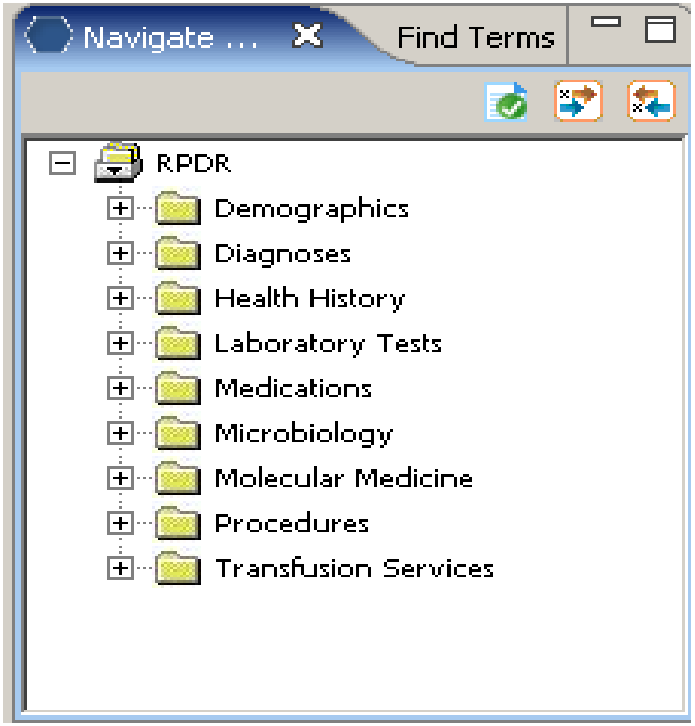
Typical i2b2 Metadata Categories

- Diagnoses
- Procedures
- Demographics
- Lab Tests
- Encounters (visits or observations)
- Providers (observers)
- Health History (physical findings and vital signs)
- Transfusion
- Microbiology

Structure of Metadata Table

METADATA	
C_HLEVEL	INT NULL
C_FULLNAME	VARCHAR(900) NULL
C_NAME	VARCHAR(2000) NULL
C_SYNONYM_CD	CHAR(1) NULL
C_VISUALATTRIBUTES	CHAR(3) NULL
C_TOTALNUM	INT NULL
C_BASECODE	VARCHAR(450) NULL
C_METADATAXML	TEXT NULL
C_FACTTABLECOLUMN	VARCHAR(50) NULL
C_TABLENAME	VARCHAR(50) NULL
C_COLUMNNAME	VARCHAR(50) NULL
C_COLUMNDATATYPE	VARCHAR(50) NULL
C_OPERATOR	VARCHAR(10) NULL
C_DIMCODE	VARCHAR(900) NULL
C_COMMENT	TEXT NULL
C_TOOLTIP	VARCHAR(900) NULL
UPDATE_DATE	DATETIME NULL
DOWNLOAD_DATE	DATETIME NULL
IMPORT_DATE	DATETIME NULL
SOURCESYSTEM_CD	VARCHAR(50) NULL
VALUETYPE_CD	VARCHAR(50) NULL

c_hlevel = 1

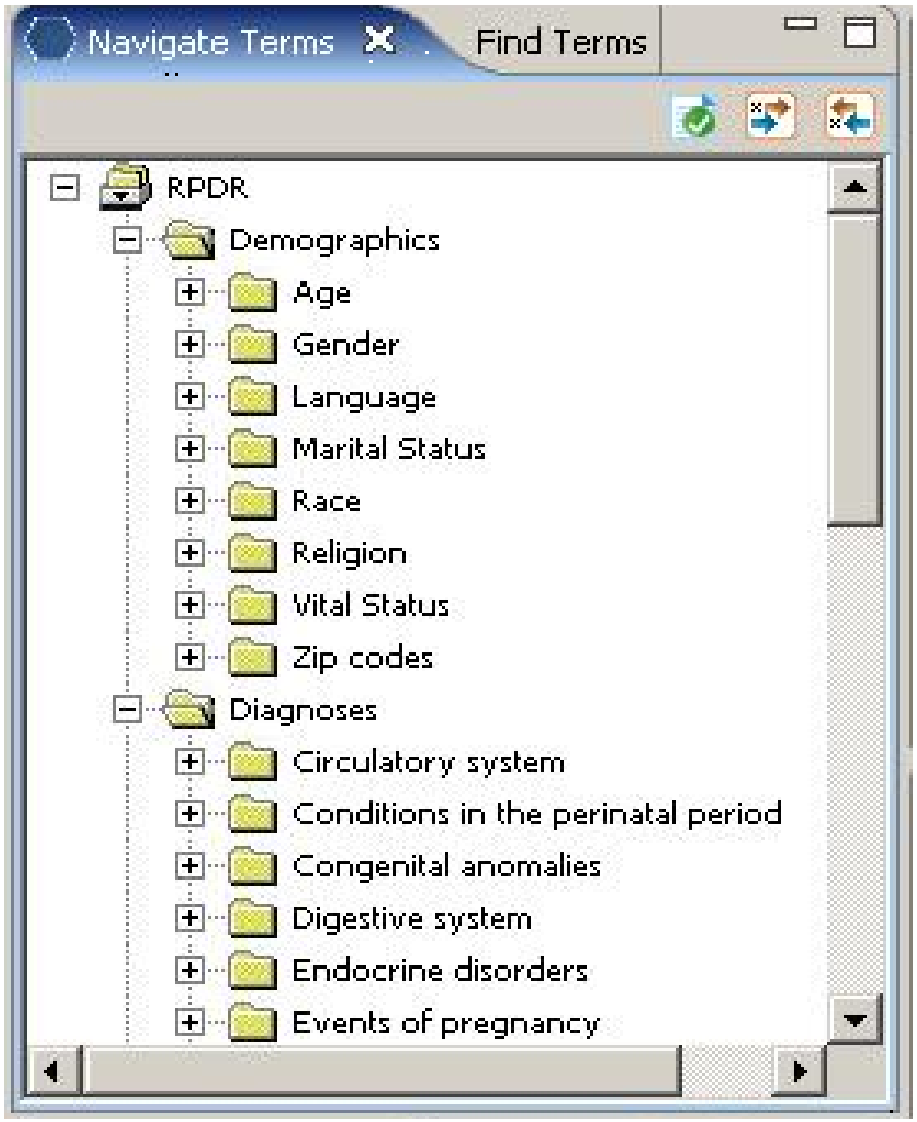



```
select c_hlevel, c_fullname,
c_name,c_visualattributes
from testrpdr
where c_hlevel=1
```

Results | Messages

	c_hlevel	c_fullname	c_name	c_visualattributes
1	1	\RPDR\Microbiology	Microbiology	FA
2	1	\RPDR\Procedures	Procedures	FA
3	1	\RPDR\Labtests	Laboratory Tests	FA
4	1	\RPDR\Medications	Medications	FA
5	1	\RPDR\Transfusions	Transfusion Services	FA
6	1	\RPDR\HealthHistory	Health History	FA
7	1	\RPDR\HPCGG	Molecular Medicine	FA
8	1	\RPDR\Diagnoses	Diagnoses	FA
9	1	\RPDR\Demographics	Demographics	FA

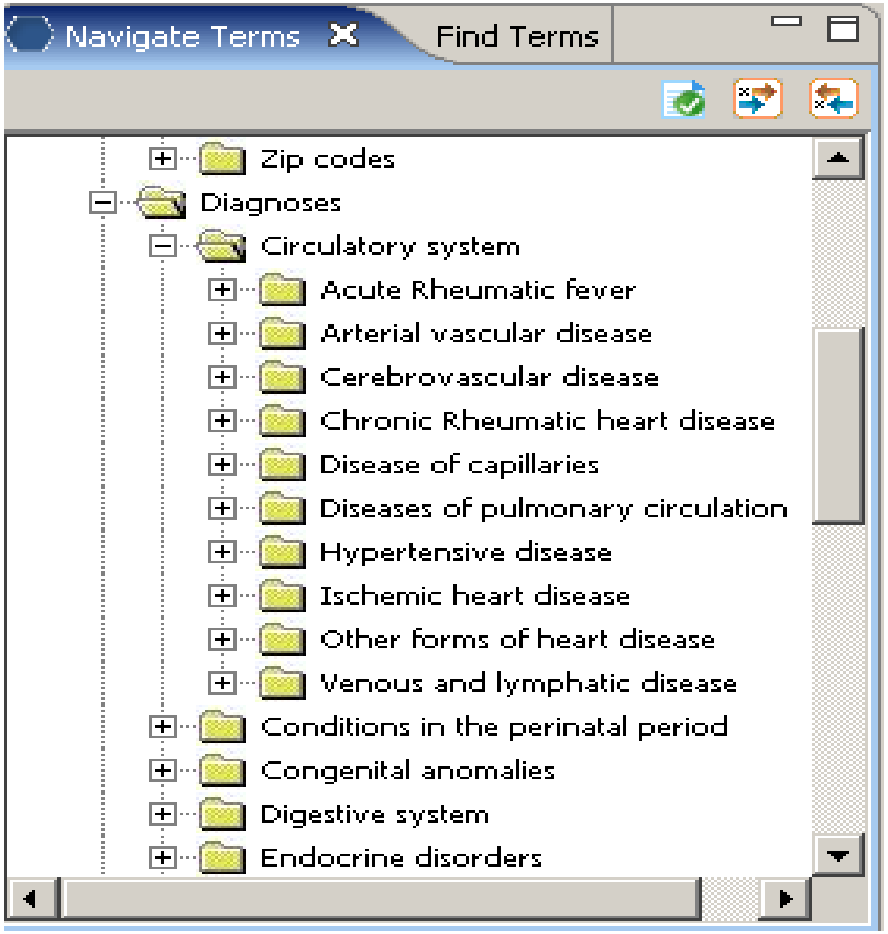
c_hlevel = 2


Navigate Terms
Find Terms
RPDR
Demographics
Age
Gender
Language
Marital Status
Race
Religion
Vital Status
Zip codes
Diagnoses
Circulatory system
Conditions in the perinatal period
Congenital anomalies
Digestive system
Endocrine disorders
Events of pregnancy
c_hlevel = 2
c_hlevel = 2

```
select c_hlevel,c_fullname,c_name,c_visualattributes
from testrpdr
where c_hlevel=2
order by c_fullname
```

C_HLEVEL	C_FULLNAME	C_NAME	C_VISUALATTRIBUTES
2	\RPDR\Demographics\Age	Age	FA
2	\RPDR\Demographics\Gender	Gender	FA
2	\RPDR\Demographics\Language	Language	FA
2	\RPDR\Demographics\Marital Status	Marital Status	FA
2	\RPDR\Demographics\Race	Race	FA
2	\RPDR\Demographics\Religion	Religion	FA
2	\RPDR\Demographics\Vital Status	Vital Status	FA
2	\RPDR\Demographics\Zip codes	Zip codes	FA
2	\RPDR\Diagnoses\Circulatory system (390-459)	Circulatory system	FA
2	\RPDR\Diagnoses\Congenital anomalies (740-759)	Congenital anomalies	FA
2	\RPDR\Diagnoses\Digestive system (520-579)	Digestive system	FA
2	\RPDR\Diagnoses\Endocrine disorders (240-259)	Endocrine disorders	FA
2	\RPDR\Diagnoses\Events of pregnancy (630-677)	Events of pregnancy	FA
2	\RPDR\Diagnoses\Genitourinary system (580-629)	Genitourinary system	FA
2	\RPDR\Diagnoses\Hematologic diseases (280-289)	Hematologic diseases	FA

c_hlevel = 3



c_hlevel=3

```
select c_hlevel, c_fullname,c_name,c_visualattributes
from testrpdr
where  c_fullname like '%Diagnoses%' and c_hlevel<4
order by c_fullname, c_hlevel
```

C_HLEVEL	C_FULLNAME	C_NAME
1	\RPDR\Diagnoses	Diagnoses
2	\RPDR\Diagnoses\Circulatory system (390-459)	Circulatory system
3	\RPDR\Diagnoses\Circulatory system (390-459)\(448) Disease of capillaries	Disease of capillaries
3	\RPDR\Diagnoses\Circulatory system (390-459)\Acute Rheumatic fever (390-392)	Acute Rheumatic fever
3	\RPDR\Diagnoses\Circulatory system (390-459)\Arterial vascular disease (440-447)	Arterial vascular disease
3	\RPDR\Diagnoses\Circulatory system (390-459)\Cerebrovascular disease (430-438)	Cerebrovascular disease
3	\RPDR\Diagnoses\Circulatory system (390-459)\Chronic Rheumatic heart disease (393-398)	Chronic Rheumatic heart dise
3	\RPDR\Diagnoses\Circulatory system (390-459)\Diseases of pulmonary circulation (415-417)	Diseases of pulmonary circul
3	\RPDR\Diagnoses\Circulatory system (390-459)\Hypertensive disease (401-405)	Hypertensive disease
3	\RPDR\Diagnoses\Circulatory system (390-459)\Ischemic heart disease (410-414)	Ischemic heart disease
3	\RPDR\Diagnoses\Circulatory system (390-459)\Other forms of heart disease (420-429)	Other forms of heart disease
3	\RPDR\Diagnoses\Circulatory system (390-459)\Venous and lymphatic disease (451-459)	Venous and lymphatic disease
2	\RPDR\Diagnoses\Conditions in the perinatal period (760-779)	Conditions in the perinatal p
3	\RPDR\Diagnoses\Conditions in the perinatal period (760-779)\(764) Slow fetal growth and fetal~	Slow fetal growth and fetal r
3	\RPDR\Diagnoses\Conditions in the perinatal period (760-779)\(765) Disorders relating to short~	Disorders relating to short g

c_fullname and c_name

METADATA	
C_HLEVEL	INT NULL
C_FULLNAME	VARCHAR(900) NULL
C_NAME	VARCHAR(2000) NULL
C_SYNONYM_CD	CHAR(1) NULL
C_VISUALATTRIBUTES	CHAR(3) NULL
C_TOTALNUM	INT NULL
C_BASECODE	VARCHAR(450) NULL
C_METADATAXML	TEXT NULL
C_FACTTABLECOLUMN	VARCHAR(50) NULL
C_TABLENAME	VARCHAR(50) NULL
C_COLUMNNAME	VARCHAR(50) NULL
C_COLUMNDATATYPE	VARCHAR(50) NULL
C_OPERATOR	VARCHAR(10) NULL
C_DIMCODE	VARCHAR(900) NULL
C_COMMENT	TEXT NULL
C_TOOLTIP	VARCHAR(900) NULL
UPDATE_DATE	DATETIME NULL
DOWNLOAD_DATE	DATETIME NULL
IMPORT_DATE	DATETIME NULL
SOURCESYSTEM_CD	VARCHAR(50) NULL
VALUETYPE_CD	VARCHAR(50) NULL

c_fullname and c_name

c_fullname is the hierarchical path that leads to the term

\RPDR
\Diagnoses
\Musculoskeletal and connective tissue (710-739)
\Arthropathies (710-719)
\(714) Rheumatoid arthritis and other arthropathies
\(714-0) Rheumatoid arthritis

c_name is the actual term

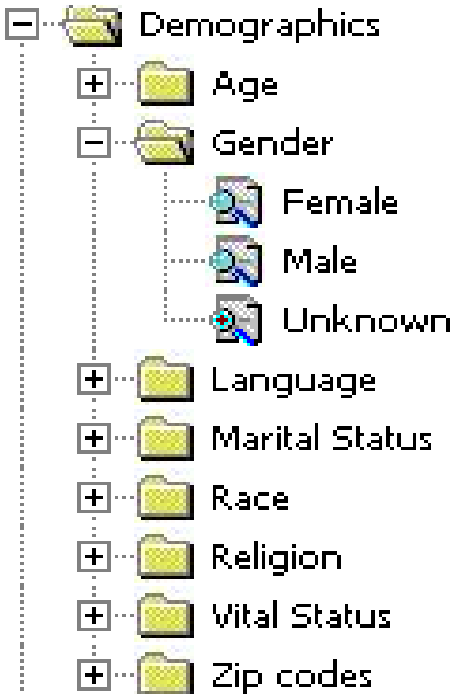
Rheumatoid arthritis
Atrophic arthritis
RA [Rheumatoid arthritis]
Chronic rheumatic arthritis

c_fullname	c_name
\RPDR\Diagnoses\Musculoskeletal and connective tissue (710-739)\Arthropathies (710-719)\(714) Rheumatoid arthritis and ot~\(714-0) Rheumatoid arthritis	Rheumatoid arthritis
\RPDR\Diagnoses\Musculoskeletal and connective tissue (710-739)\Arthropathies (710-719)\(714) Rheumatoid arthritis and ot~\(714-0) Rheumatoid arthritis	Atrophic arthritis
\RPDR\Diagnoses\Musculoskeletal and connective tissue (710-739)\Arthropathies (710-719)\(714) Rheumatoid arthritis and ot~\(714-0) Rheumatoid arthritis	RA [Rheumatoid arthritis]
\RPDR\Diagnoses\Musculoskeletal and connective tissue (710-739)\Arthropathies (710-719)\(714) Rheumatoid arthritis and ot~\(714-0) Rheumatoid arthritis	Chronic rheumatic arthritis

c_synonym_cd

C_HLEVEL	C_NAME	C_SYNONYM_CD	C_VISUALATTRIBUTES	C_BASECODE
5	Other rheumatoid arthritis with visceral or systemic involvement	N	LA	7142
5	Rheumatoid carditis	Y	LH	7142
5	Rheumatoid arthritis with other visceral or systemic involvement	Y	LH	7142
5	Enteropathic arthritis	Y	LH	7142
5	Poulet's disease	Y	LH	7142

c_visualattributes



1st character :

F = Folder

C = Container

M = Multiple

2nd character:

A = Active

I = Inactive

H = Hidden

C_HLEVEL	C_NAME	C_SYNONYM_CD	C_VISUALATTRIBUTES	C_BASECODE	
5	Other specified inflammatory polyarthropathies	N	FA	7148	(non-specific code)
5	Other specified inflammatory polyarthropathies	Y	FH	7148	
6	Other specified inflammatory polyarthropathies	N	LA	71489	(specific code)

c_basecode

The basecode is the coded value for the term.

c_hlevel	c_fullname	c_name	c_visualattributes	c_basecode	c_operator
2	\RPDR\Demographics\Gender	Gender	FA	NULL	LIKE
3	\RPDR\Demographics\Gender\Female	Female	LA	DEM|SEX:f	LIKE
3	\RPDR\Demographics\Gender\Male	Male	LA	DEM|SEX:m	LIKE
3	\RPDR\Demographics\Gender\Unknown	Unknown	MA	NULL	LIKE
4	\RPDR\Demographics\Gender\Unknown\Unknown-@	Unknown-@	LH	DEM|SEX:@	LIKE
4	\RPDR\Demographics\Gender\Unknown\Unknown-U	Unknown-U	LH	DEM|SEX:u	LIKE

It maps to the concept_cd in the star schema tables.

c_facttablecolumn,c_tablename,c_columnname, c_columndatatype,c_operator,c_dimcode

METADATA	
C_HLEVEL	INT NULL
C_FULLNAME	VARCHAR(900) NULL
C_NAME	VARCHAR(2000) NULL
C_SYNONYM_CD	CHAR(1) NULL
C_VISUALATTRIBUTES	CHAR(3) NULL
C_TOTALNUM	INT NULL
C_BASECODE	VARCHAR(450) NULL
C_METADATAXML	TEXT NULL
C_FACTTABLECOLUMN	VARCHAR(50) NULL
C_TABLENAME	VARCHAR(50) NULL
C_COLUMNNAME	VARCHAR(50) NULL
C_COLUMNDATATYPE	VARCHAR(50) NULL
C_OPERATOR	VARCHAR(10) NULL
C_DIMCODE	VARCHAR(900) NULL
C_COMMENT	TEXT NULL
C_TOOLTIP	VARCHAR(900) NULL
UPDATE_DATE	DATETIME NULL
DOWNLOAD_DATE	DATETIME NULL
IMPORT_DATE	DATETIME NULL
SOURCESYSTEM_CD	VARCHAR(50) NULL
VALUETYPE_CD	VARCHAR(50) NULL

Fields used to construct queries

c_facttablecolumn	c_tablename	c_columnname	c_columndatatype	c_operator	c_dimcode
concept_cd	concept_dimension	concept_path	T	LIKE	\RPDR\Demographics\Gender
concept_cd	concept_dimension	concept_path	T	LIKE	\RPDR\Demographics\Gender\Female
concept_cd	concept_dimension	concept_path	T	LIKE	\RPDR\Demographics\Gender\Male
concept_cd	concept_dimension	concept_path	T	LIKE	\RPDR\Demographics\Gender\Unknown
concept_cd	concept_dimension	concept_path	T	LIKE	\RPDR\Demographics\Gender\Unknown\Unknown-@
concept_cd	concept_dimension	concept_path	T	LIKE	\RPDR\Demographics\Gender\Unknown\Unknown-U

Select * from observation_fact where c_facttablecolumn in
(select concept_cd from c_tablename where c_columname c_operator ‘c_dimcode%’)

Select * from observation_fact where concept_cd in
(select concept_cd from concept_dimension where concept_path like
’\RPDR\Demographics\Gender%’)

c_metadataxml

stores values and information about the concept such as high and low indicators

concept_dimension table

concept_dimension		
PK	**concept_path**	**VARCHAR(700)**
	concept_cd name_char concept_blob	**VARCHAR(20)** VARCHAR(2000) TEXT(10)

Outline of Workshop #1

- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

Navigating concept_dimension

Except for diagnoses, which do not have a standard prefix, the prefix of the location_path identifies the type of concept:

LAB\ is for Laboratory data
TRS\ is for Transfusion data
MIC\ is for Microbiology data
MCP\ is for CPT medications
MUL\ is for hospital medications
PRC\ is for procedures

So, if looking for a list of all possible labs, the query to run is

select * from concept_dimension where concept_path like 'lab\%'

concept_cd can be joined to concept_cd in the Observation_Fact table.

Visit_dimension

- Encounter number
- In/out code
- Hospital of service
- Clinic

Encounter_num can be joined to encounter_num in the Observation_Fact table.

Patient_num can be joined to patient_num in the Observation_Fact and Visit_Dimension tables.

Example 1: Query based on a diagnosis

To find all the patients diagnosed with migraines, use this query:

```
Select p.patient_num
From patient_dimension p join observation_fact f
On p.patient_num = f.patient_num
And f.concept_cd in
(select concept_cd
from concept_dimension
where concept_path like
'Neurologic Disorders (320-389)\(346) Migraine\%')
```

Add Demographics

To find the ages of all patients diagnosed with migraines, use this query:

```
Select p.age_in_years_num
From patient_dimension p, observation_fact f
Where p.patient_num = f.patient_num
And f.concept_cd in
(select concept_cd
from concept_dimension
where concept_path like
'Neurologic Disorders (320-389)\(346) Migraine\%')
```

Add Visit data

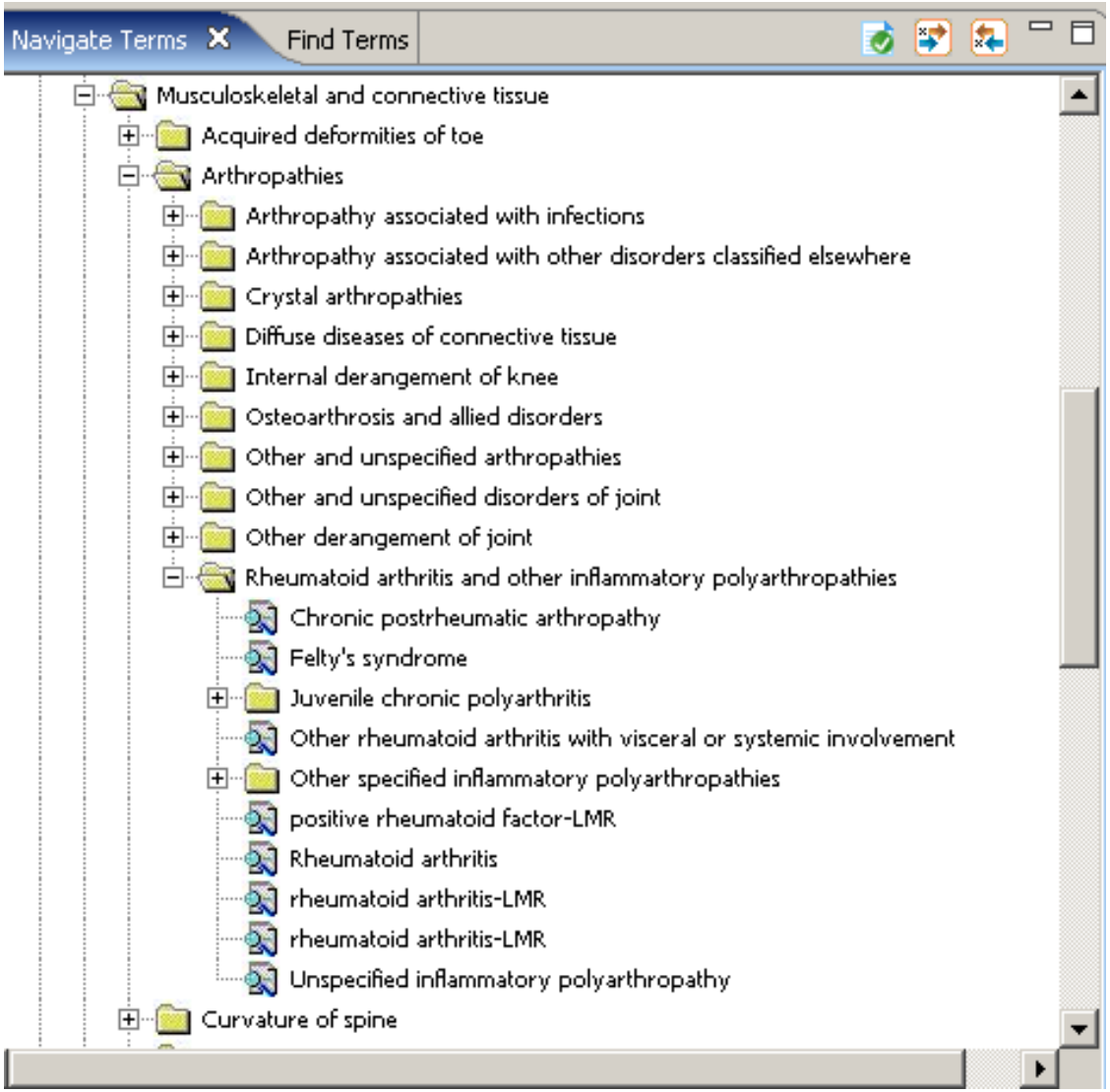
To limit the query to only those patients seen at MGH, add the Visit table.

```
Select p.patient_num,p.age_in_years_num
From patient_dimension p join observation_fact f
on p.patient_num = f.patient_num join visit_dimension v
on v.encounter_num = f.encounter_num
And v.location_cd = 'MGH'
And f.concept_cd in
(select concept_cd
from concept_dimension
where concept_path like
'Neurologic Disorders (320-389)\(346) Migraine\%')
```

Example 2: A more complex query

Find patients with Rheumatoid Arthritis

icd9 714.xx



```
select distinct(concept_cd),name_char
from concept_dimension
where concept_path like '%rheumatoid%'
```

714	Rheumatoid arthritis and other inflammatory polyarthropathies
7140	Rheumatoid arthritis
7141	Felty's syndrome
7142	Other rheumatoid arthritis with visceral or systemic involvement
7143	Juvenile chronic polyarthritis
71430	Polyarticular juvenile rheumatoid arthritis, chronic or unspecified
71431	Polyarticular juvenile rheumatoid arthritis, acute
71432	Pauciarticular juvenile rheumatoid arthritis
71433	Monoarticular juvenile rheumatoid arthritis
7144	Chronic postrheumatic arthropathy
7148	Other specified inflammatory polyarthropathies
71481	Rheumatoid lung
71489	Other specified inflammatory polyarthropathies
7149	Unspecified inflammatory polyarthropathy
C86430	Rheumatoid factor; qualitative
C86431	Rheumatoid factor; quantitative
LPA344	Positive rheumatoid factor-LMR 344
LPA377	Rheumatoid arthritis-LMR 377
LPA404	Stills disease-LMR 404
LPA924	Juvenile polyarthritis-LMR 924
LPB2254	Rheumatoid arthritis-LMR 2254
V821	Screening for rheumatoid arthritis

Get the number of visits with RA as dx over time for each patient, excluding Juvenile RA 714.30

```
Select patient_num, count(*) as RAvisits,
min(start_date) as startdate,max(start_date) as enddate,
datediff(month,min(start_date),max(start_date)) as RAmonthsobserved
Into RAvisitsbypatient
from observation_fact
where (concept_cd like '714%'
or concept_cd like 'lp%2254' or concept_cd like 'lp%377')
and concept_cd <>'714.30'
group by patient_num
```

RAvisits	startdate	enddate	RAmonthsobserved
15	2006-09-13 00:00:00.000	2007-05-29 00:00:00.000	8
15	2006-10-17 00:00:00.000	2007-06-22 00:00:00.000	8
15	2006-08-30 00:00:00.000	2007-05-23 00:00:00.000	9
15	1997-07-30 00:00:00.000	1998-04-07 00:00:00.000	9
15	2005-12-05 00:00:00.000	2006-11-30 00:00:00.000	11
15	2006-06-06 00:00:00.000	2007-05-24 00:00:00.000	11
15	2006-09-22 00:00:00.000	2007-08-06 00:00:00.000	11
15	2006-07-17 00:00:00.000	2007-06-19 00:00:00.000	11
15	2002-01-02 00:00:00.000	2002-12-31 00:00:00.000	11

Multiple paths

```
select distinct(patient_num)  into BoneMarrowTransplants from observation_fact where
concept_cd in
(Select concept_cd
From concept_dimension
Where concept_path LIKE
'PRC\ICD9 (Inpatient)\(40-41) Operations on hemic and lymphatic system\(p41) Operations on
      bone marrow a~\(p41-0) Bone marrow or hematopoie~\%'
or concept_path LIKE 'PRC\CPT\(10021-69990) Surgery\(38100-38999) Hemic and Lymphatic
      Systems\(38204-38242) Bone Marrow or Stem Cell\(38242) Bone marrow or blood-deri~\%'
or concept_path LIKE 'PRC\CPT\(10021-69990) Surgery\(38100-38999) Hemic and Lymphatic
      Systems\(38204-38242) Bone Marrow or Stem Cell\(38240) Bone marrow or blood-deri~\%'
or concept_path LIKE 'PRC\CPT\(10021-69990) Surgery\(38100-38999) Hemic and Lymphatic
      Systems\(38204-38242) Bone Marrow or Stem Cell\(38241) Bone marrow or blood-deri~\%'
or concept_path LIKE '(Pre) Transplants and Tracheostomy\Surgical\(481) Bone Marrow
      Transplant\%'
or concept_path LIKE 'zz V-codes\Conditions influencing health status (V40-V49)\(V42) Organ or
      tissue replaced by~\(V42-8) Other specified organ or ~\(V42-81) Bone marrow replaced by ~\%'
or concept_path LIKE 'PRC\LMR\(LPA547) bone marrow transplant\%'
or concept_path LIKE 'Injury and poisoning (800-999)\Complications of medical care (996-
      999)\(996) Complications peculiar to c~\(996-8) Complications of transpla~\(996-85)
      Complications of bone ma~\%')
```

Find CCP lab codes

```
select c_hlevel,c_fullname,c_name,c_basecode into CCPCodes
from labtests where c_fullname like '%ACCP%'
or c_fullname like '%ANTCCP%'
```

c_hlevel	c_fullname	c_name	c_basecode
3	LAB\(LLB63) DBNull\(LLB64) DBNull\ACCP	CCP Ab (Group:ACCP)	ACCP
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ACCP\BC1-1318	A-CCP (Test:bc1-1318)	BC1-1318
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ACCP\WC890.0334	CYCLIC CITRULLINATED PEPTIDE (Test:wc890.0334)	WC890.0334
3	LAB\(LLB63) DBNull\(LLB64) DBNull\ANTCCP	CCP Ab, IgG (Group:ANTCCP)	ANTCCP
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ANTCCP\FC50.05	ANTI CYCLIC CITRULLINE PEPTIDE (Test:fc50.05)	FC50.05
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ANTCCP\LC6138	CCP IgG Antibodies (Test:lc6138)	LC6138
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ANTCCP\MCSQ-AN...	Anti CCP (Test:mcsq-antccp)	MCSQ-ANTCCP
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ANTCCP\MISQ-CCPT	anti-CCP IgG (Test:misq-ccpt)	MISQ-CCPT
4	LAB\(LLB63) DBNull\(LLB64) DBNull\ANTCCP\NCCCPG	ANTI CCP IGG (Test:ncccpg)	NCCCPG

Determine which patients have a CCP titer >40 and have been seen in either
the BWH Arthritis center or MGH Arthritis Associates.

```
select distinct(patient_num)
from observation_fact
where concept_cd in
(select c_basecode from CCPcodes) and nval_num >40
and patient_num in
(select patient_num from vg_ravisitsbypatient)
and patient_num in
(select distinct(patient_num)
from visit_dimension
where location_path like '%arthritis%')
```

Tips

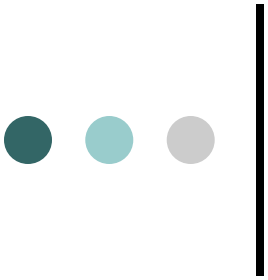
- Look at the tables
- Bear in mind that you won't need to concentrate on every field in every table, but can drill down into the particular fields of interest as needed.
- Figure out how the dimension tables tie into the fact table.
- Check out the mapping tables that are included if you need to further identify the data.
- Try running one of the sample queries.
- Think about what questions you want answered,then try to frame them based on the data in the data mart.
- Write the SQL to perform your queries. Start slowly and gradually build up the complexity of each query.

Outline of Workshop #1

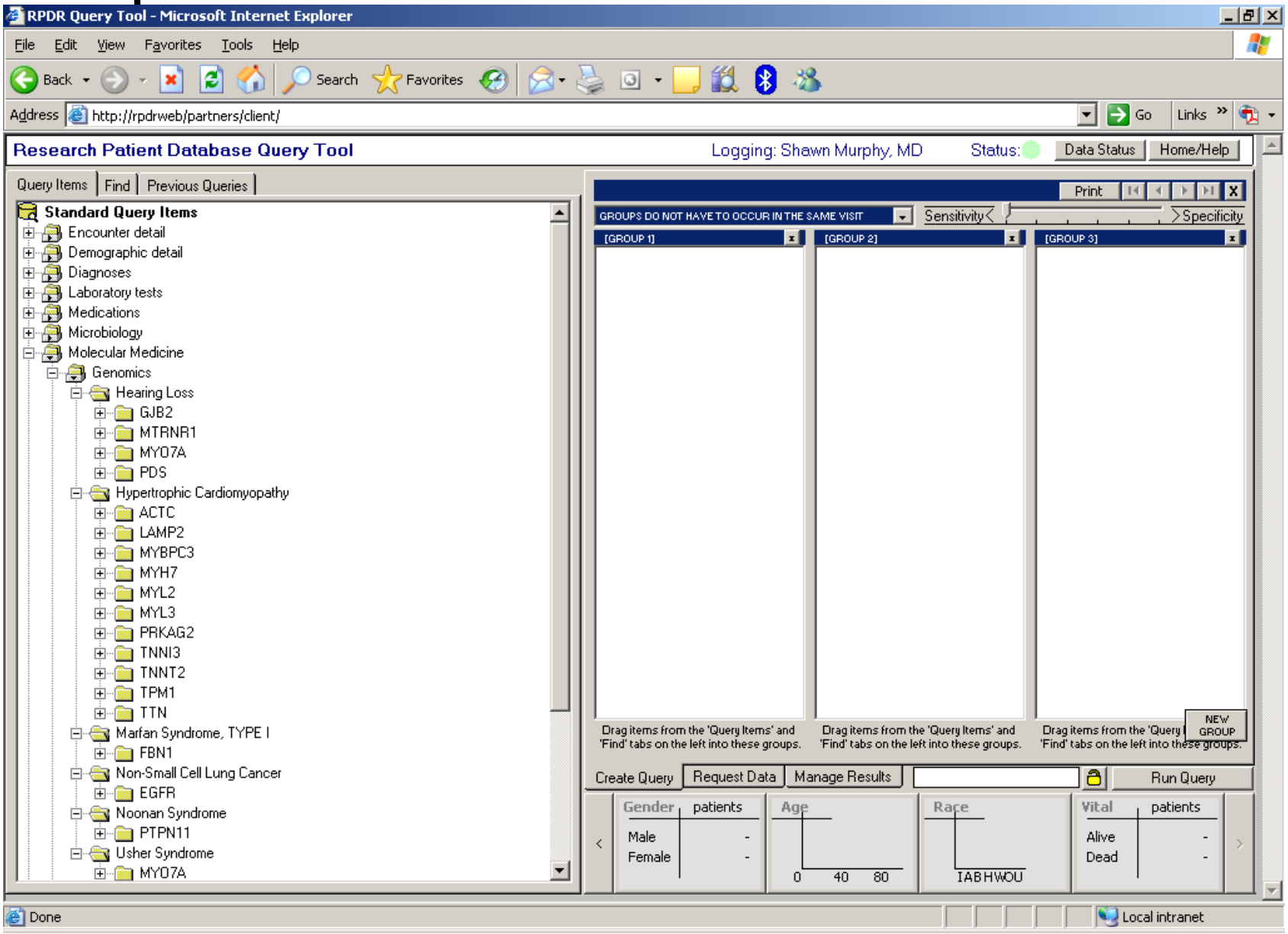
- What data is most useful
- Data commonly available
- Privacy of clinical data
- Data models
- Querying the clinical data
- Transformations necessary to load genomic data

Transforming Genomic Data

- SNP / Variants
- RNA Expression Values (GEO)



- Variant transform method



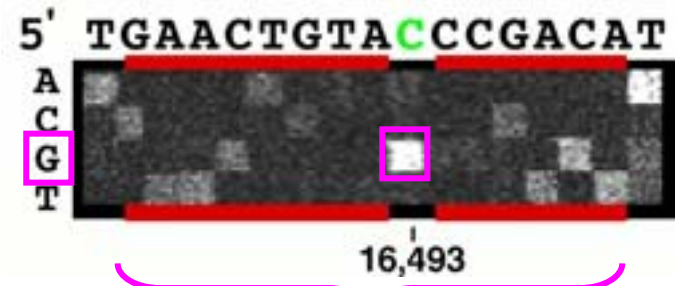
Antonarakis Variant Notation

Wildtype Sequence

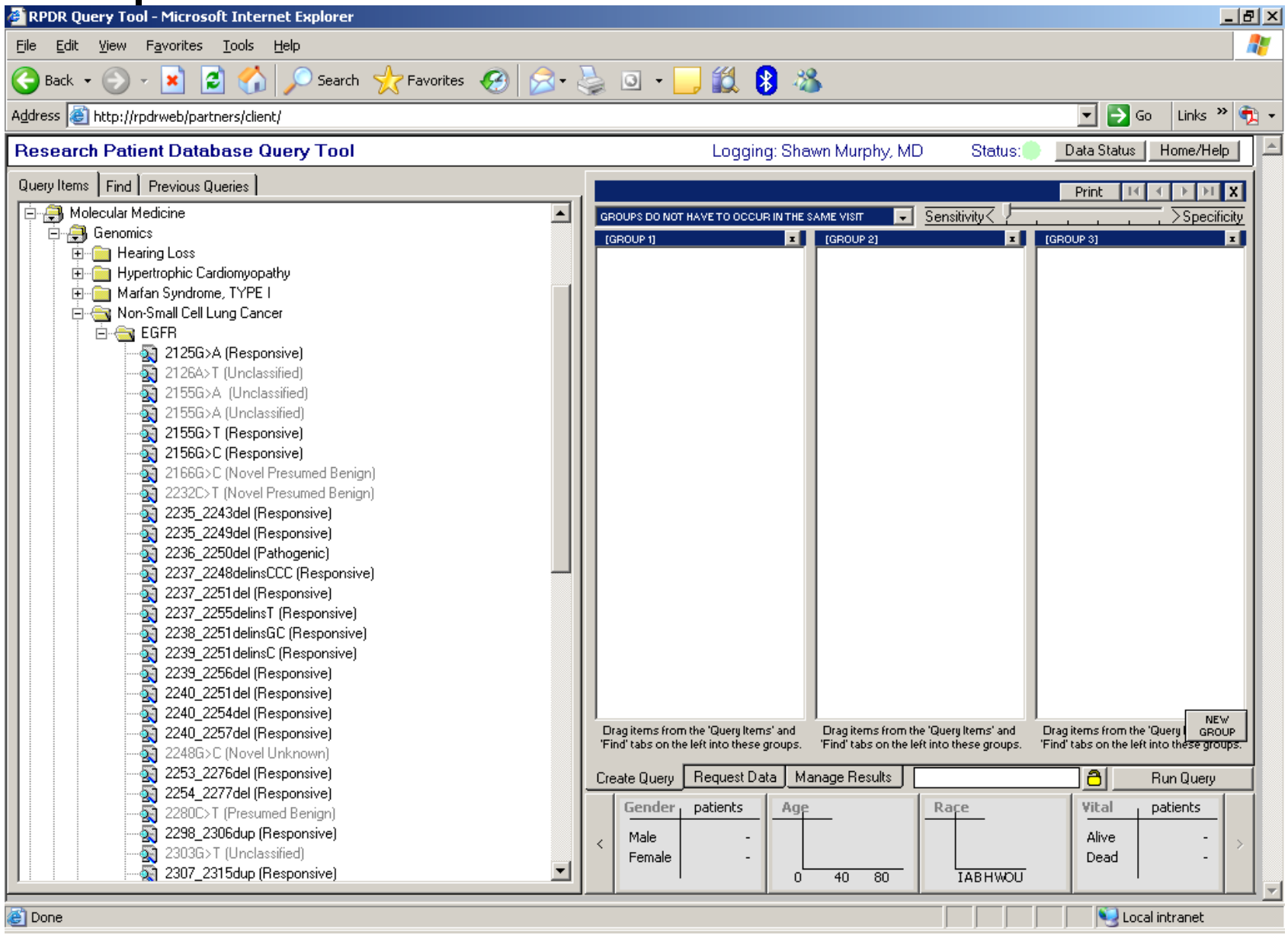
```
5' ..TGAACTGTATCCGACAT..
3'      tgacatAggctgtag
        tgacatCggctgtag
        tgacatGggctgtag
        tgacatTggctgtag
3'       gacataAgctgtaga
         gacataCgctgtaga
         gacataGgctgtaga
         gacataTgctgtaga
```

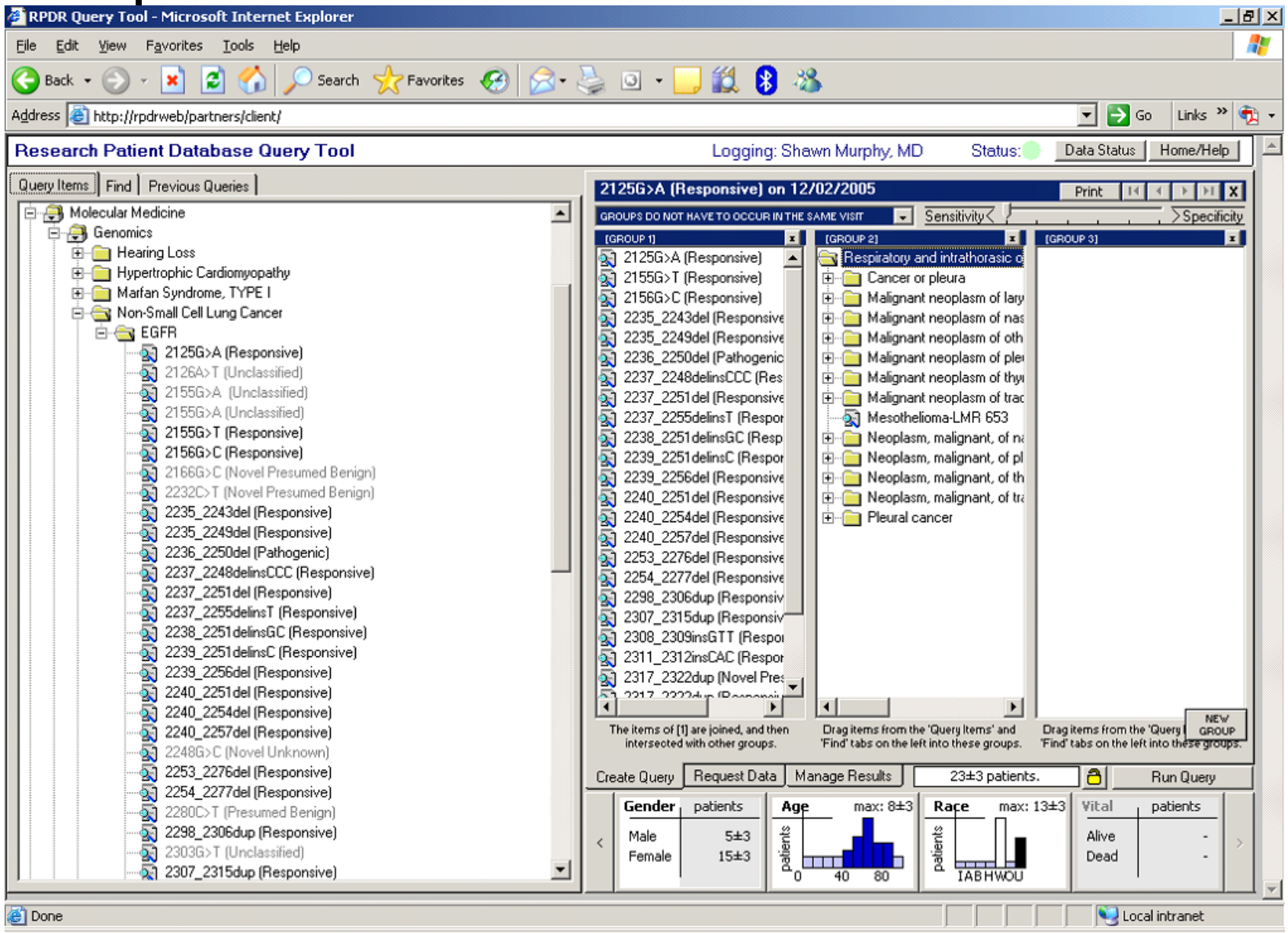
Variant

```
5' ..TGAACTGTACCCGACAT..
3'      tgacatAggctgtag
        tgacatCggctgtag
        tgacatGggctgtag
        tgacatTggctgtag
3'       gacataAgctgtaga
         gacataCgctgtaga
         gacataGgctgtaga
         gacataTgctgtaga
```

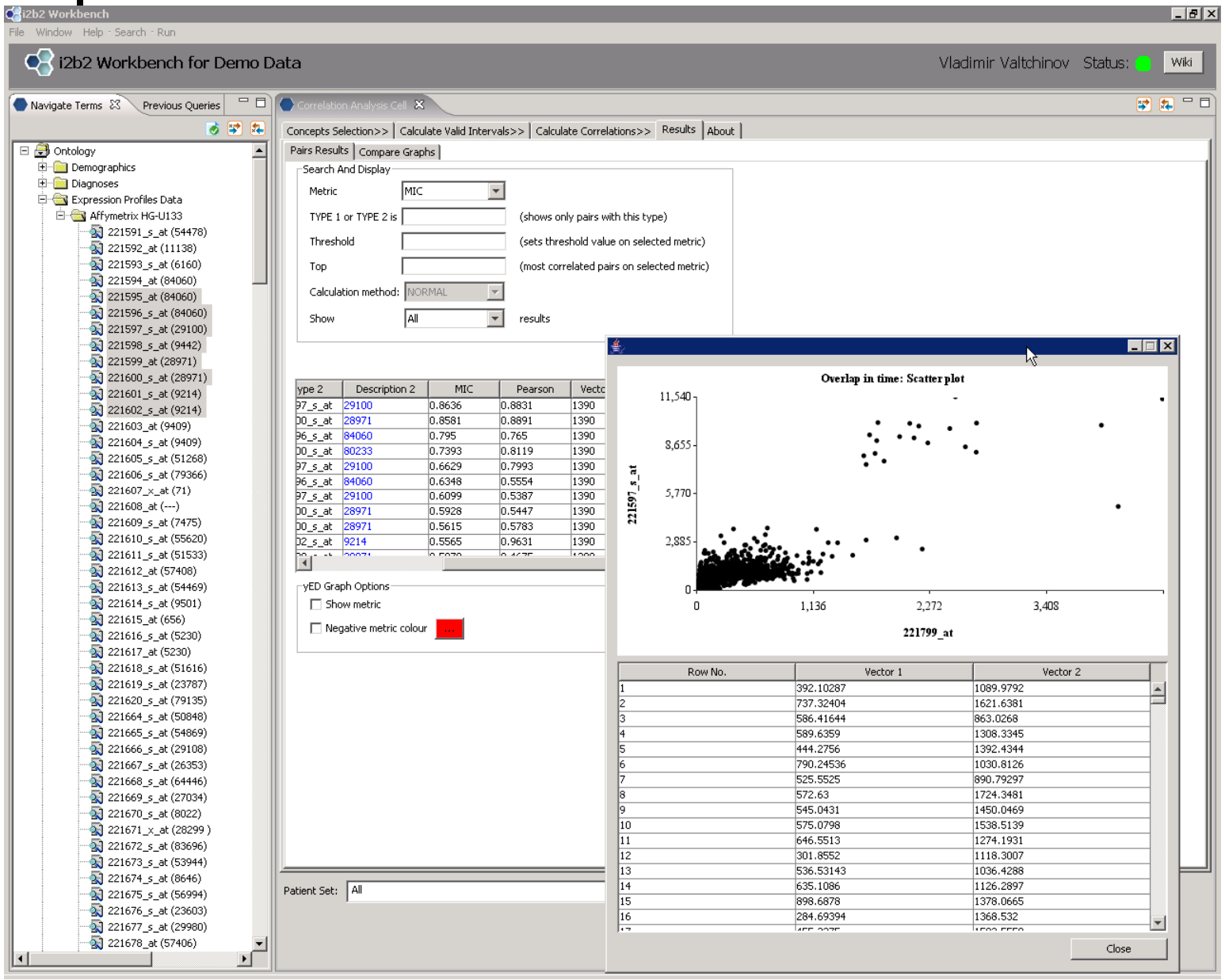


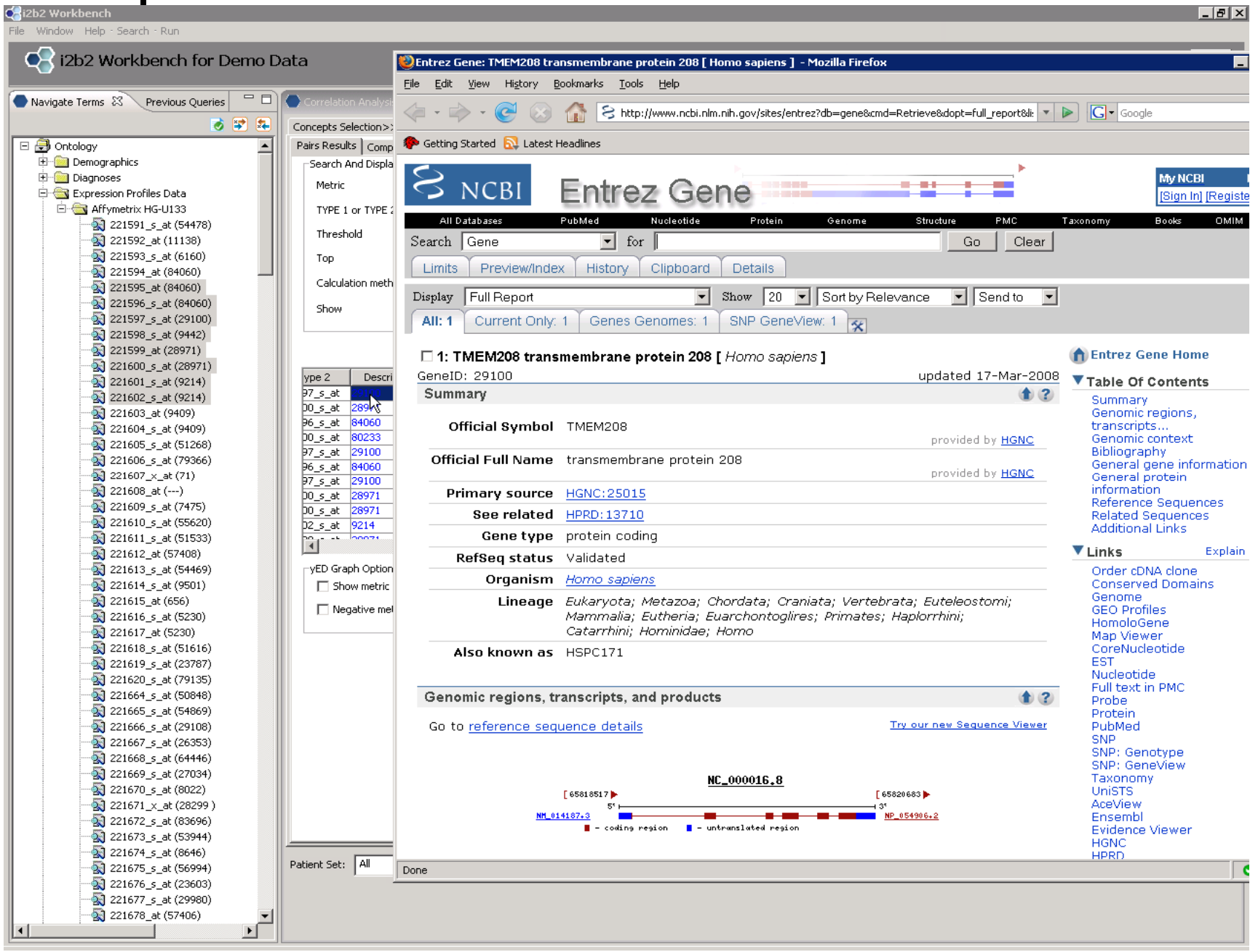
Footprint





- GEO transform method


i2b2 Workbench
File Window Help · Search · Run
i2b2 Workbench for Demo Data
Vladimir Valtchinov Status:
Wiki
Navigate Terms
Previous Queries
Ontology
Demographics
Diagnoses
Expression Profiles Data
Affymetrix HG-U133
221591_s_at (54478)
221592_at (11138)
221593_s_at (6160)
221594_at (84060)
221595_at (84060)
221596_s_at (84060)
221597_s_at (29100)
221598_s_at (9442)
221599_at (28971)
221600_s_at (28971)
221601_s_at (9214)
221602_s_at (9214)
221603_at (9409)
221604_s_at (9409)
221605_s_at (51268)
221606_s_at (79366)
221607_x_at (71)
221608_at (---)
221609_s_at (7475)
221610_s_at (55620)
221611_s_at (51533)
221612_at (57408)
221613_s_at (54469)
221614_s_at (9501)
221615_at (656)
221616_s_at (5230)
221617_at (5230)
221618_s_at (51616)
221619_s_at (23787)
221620_s_at (79135)
221664_s_at (50848)
221665_s_at (54869)
221666_s_at (29108)
221667_s_at (26353)
221668_s_at (64446)
221669_s_at (27034)
221670_s_at (8022)
221671_x_at (28299)
221672_s_at (83696)
221673_s_at (53944)
221674_s_at (8646)
221675_s_at (56994)
221676_s_at (23603)
221677_s_at (29980)
221678_at (57406)
Correlation Analysis Cell
Concepts Selection>>
Calculate Valid Intervals>>
Calculate Correlations>>
Results
About
Pairs Results
Compare Graphs
Search And Display
Metric
MIC
TYPE 1 or TYPE 2 is
(shows only pairs with this type)
Threshold
(sets threshold value on selected metric)
Top
(most correlated pairs on selected metric)
Calculation method:
NORMAL
Show
All
results
ype 2
Description 2
MIC
Pearson
Vecto
97_s_at 29100 0.8636 0.8831 1390
00_s_at 28971 0.8581 0.8891 1390
96_s_at 84060 0.795 0.765 1390
00_s_at 80233 0.7393 0.8119 1390
97_s_at 29100 0.6629 0.7993 1390
96_s_at 84060 0.6348 0.5554 1390
97_s_at 29100 0.6099 0.5387 1390
00_s_at 28971 0.5928 0.5447 1390
00_s_at 28971 0.5615 0.5783 1390
02_s_at 9214 0.5565 0.9631 1390
yED Graph Options
Show metric
Negative metric colour
Patient Set:
All
Overlap in time: Scatter plot
11,540
8,655
5,770
2,885
0
221597_s_at
0
1,136
2,272
3,408
221799_at
Row No.
Vector 1
Vector 2
1 392.10287 1089.9792
2 737.32404 1621.6381
3 586.41644 863.0268
4 589.6359 1308.3345
5 444.2756 1392.4344
6 790.24536 1030.8126
7 525.5525 890.79297
8 572.63 1724.3481
9 545.0431 1450.0469
10 575.0798 1538.5139
11 646.5513 1274.1931
12 301.8552 1118.3007
13 536.53143 1036.4288
14 635.1086 1126.2897
15 898.6878 1378.0665
16 284.69394 1368.532
Close


i2b2 Workbench
File Window Help - Search - Run
i2b2 Workbench for Demo Data
Navigate Terms
Previous Queries
Ontology
Demographics
Diagnoses
Expression Profiles Data
Affymetrix HG-U133
221591_s_at (54478)
221592_at (11138)
221593_s_at (6160)
221594_at (84060)
221595_at (84060)
221596_s_at (84060)
221597_s_at (29100)
221598_s_at (9442)
221599_at (28971)
221600_s_at (28971)
221601_s_at (9214)
221602_s_at (9214)
221603_at (9409)
221604_s_at (9409)
221605_s_at (51268)
221606_s_at (79366)
221607_x_at (71)
221608_at (---)
221609_s_at (7475)
221610_s_at (55620)
221611_s_at (51533)
221612_at (57408)
221613_s_at (54469)
221614_s_at (9501)
221615_at (656)
221616_s_at (5230)
221617_at (5230)
221618_s_at (51616)
221619_s_at (23787)
221620_s_at (79135)
221664_s_at (50848)
221665_s_at (54869)
221666_s_at (29108)
221667_s_at (26353)
221668_s_at (64446)
221669_s_at (27034)
221670_s_at (8022)
221671_x_at (28299)
221672_s_at (83696)
221673_s_at (53944)
221674_s_at (8646)
221675_s_at (56994)
221676_s_at (23603)
221677_s_at (29980)
221678_at (57406)
Correlation Analysis
Concepts Selection>>
Pairs Results
Comp
Search And Displa
Metric
TYPE 1 or TYPE 2
Threshold
Top
Calculation meth
Show
ype 2
Descri
97_s_at
00_s_at
289
96_s_at
84060
00_s_at
80233
97_s_at
29100
96_s_at
84060
97_s_at
29100
00_s_at
28971
00_s_at
28971
02_s_at
9214
yED Graph Option
Show metric
Negative mel
Patient Set:
All
Entrez Gene: TMEM208 transmembrane protein 208 [Homo sapiens] - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=Retrieve&dopt=full_report&li
Google
Getting Started
Latest Headlines
NCBI
Entrez Gene
My NCBI
[Sign In] [Registe
All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM
Search Gene for Go Clear
Limits Preview/Index History Clipboard Details
Display Full Report Show 20 Sort by Relevance Send to
All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1
1: TMEM208 transmembrane protein 208 [Homo sapiens]
GeneID: 29100
updated 17-Mar-2008
Summary
Official Symbol TMEM208
provided by HGNC
Official Full Name transmembrane protein 208
provided by HGNC
Primary source HGNC:25015
See related HPRD:13710
Gene type protein coding
RefSeq status Validated
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as HSPC171
Genomic regions, transcripts, and products
Go to reference sequence details
Try our new Sequence Viewer
NC_000016.8
[65818517
[65820683
5'
3'
NM_014187.3
NP_054906.2
- coding region
- untranslated region
Entrez Gene Home
Table Of Contents
Summary
Genomic regions, transcripts...
Genomic context
Bibliography
General gene information
General protein information
Reference Sequences
Related Sequences
Additional Links
Links
Explain
Order cDNA clone
Conserved Domains
Genome
GEO Profiles
HomoloGene
Map Viewer
CoreNucleotide
EST
Nucleotide
Full text in PMC
Probe
Protein
PubMed
SNP
SNP: Genotype
SNP: GeneView
Taxonomy
UniSTS
AceView
Ensembl
Evidence Viewer
HGNC
HPRD
Done

References

- Murphy, S. N., Morgan, M. M., Barnett, G. O., Chueh, H. C. Optimizing Healthcare Research Data Warehouse Design through Past COSTAR Query Analysis. Proc AMIA Symp. 1999;892-96.
- Nigrin, D.J., Kohane, I.S. Data Mining by Clinicians. Proc AMIA Symp. 1998;957-61.
- Banhart, F., Klaeren, H. A Graphical Query Generator for Clinical Research Databases. Meth Inform Med 1995; 34:328-39.
- Kimball, R. The Data Warehousing Toolkit. New York: John Wiley, 1997.
- Inmon, W.H. Building the Data Warehouse, 2nd Edition. . New York: John Wiley, 1996.
- Nadkarni, P.M., Brandt, C. Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database. J Am Med Inform Assoc. 1998; 5:511-7.
- Safran, C., Porter, D., Lightfoot, J. et. al. ClinQuery: A system for online searching of data in a teaching hospital. Ann Intern Med. 1989; 111(9):751-56.

References

- Rules and Regulations. Federal Register 65(250), Section 164.514, 82818, Dec 28, 2000.
- Fischetti M,Salazar J.Model and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control. Mathematical Programming 1999; 84:283-312.
- Ohno-Machado, L., Dreiseitl,S., Vinterbo, S., Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modeling Performance, Proc AMIA Fall Symp 2001; 503-7.
- Dreiseitl,S., Vinterbo, S., Ohno-Machado, L., Disambiguation Data: Extracting Information from Anonymized Sources, Proc AMIA Fall Symp 2001; 144-8.
- J. E. Gentle, Random Number Generation and Monte Carlo Methods (Statistics and Computing), Springer-Verlag, 1998
- Sweeney L. Guaranteeing anonymity when sharing medical data, the DataFly system. Proc. AMIA Fall Symposium. 1997; 51-5.
- Murphy, S.N., Chueh, H (2002). A Security Architecture for Query Tools Used to Access Large Biomedical Databases. AMIA, Fall Symp. 2002, pages 552-556.

References

- Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D., Shneiderman, B. LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records, AMIA, Fall Symposium 1998, pp. 76-8.
- Murphy, S.N., Gainer, V.S., Chueh, H. A Visual Interface Designed for Novice Users to find Research Patient Cohorts in a Large Biomedical Database. AMIA, Fall Symp. 2003: 489-493.
- Hobbs JR. Information extraction from biomedical text. Journal of Biomedical Informatics 2002;35(4):260-4.
- Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association 1994;1(2):161-74.
- Krauthammer M, Hripsack G. A knowledge model for the interpretation and visualization of NLP-parsed discharge summaries. In: Suzanne Bakken RD, editor. Annual Symposium of the American Medical Information Association; 2001; Washington, DC: Hanley & Belfus, Inc.; 2001. p. 339-43.

References

- Xu H, Friedman C. Facilitating Research in Pathology using Natural Language Processing. In: Annual Symposium of the American Medical Informatics Association; 2003; Washington, DC: Hanley & Belfus, Inc.; 2003. p. 1057.
- Zingmond D, Lenert L. Monitoring free text data using medical language processing. Comput Biomed Res 1993;26:467-81.
- Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. Journal of Biomedical Informatics 2002;35(4):236-46.
- Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical phrase identification system (CAPIS). In: Fifteenth Annual Symposium on Computer Applications in Medical Care; 1991; Washington, DC: McGraw-Hill, Inc.; 1991. p. 843-7.
- Mikkelsen G, Aasly J. Manual semantic tagging to improve access to information in narrative electronic medical records. International Journal of Medical Informatics 2002;65:17-29.
- Brown PJB, Sonksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient databse using a semantic terminological model. Journal of the American Medical Informatics Association 2000;7:392-403.

References

- Jerry Alan Fails, Amy Karlson, Layla Shahamat, Ben Shneiderman; A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories, VAST Symposia 2006
- Barrows RC, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. In: J. Marc Overhage MP, editor. Annual Symposium of the American Medical Informatics Association; 2000; Los Angeles, CA: Hanley & Belfus, Inc.; 2000. p. 51-5.
- Johnson SB, Friedman C. Integrating Data from Natural Language Processing into a Clinical Information System. In: Cimino JJ, editor. Annual Symposium of the American Medical Informatics Association; 1996; Washington, DC: Hanley & Belfus, Inc.; 1996. p. 537-41.
- Singh, J.A., A.R. Holmgren, and S. Noorbaloochi, *Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis.* Arthritis Rheum. 2004 Dec 15;51(6):952-7.
- Valinsky LJ, Hockey RL, Hobbs MS, Fletcher DR, Pikora TJ, Parsons RW, Tan P Finding bile duct injuries using record linkage: a validated study of complications following cholecystectomy. J Clin Epidemiol 1999 Sep;52(9):893-901.