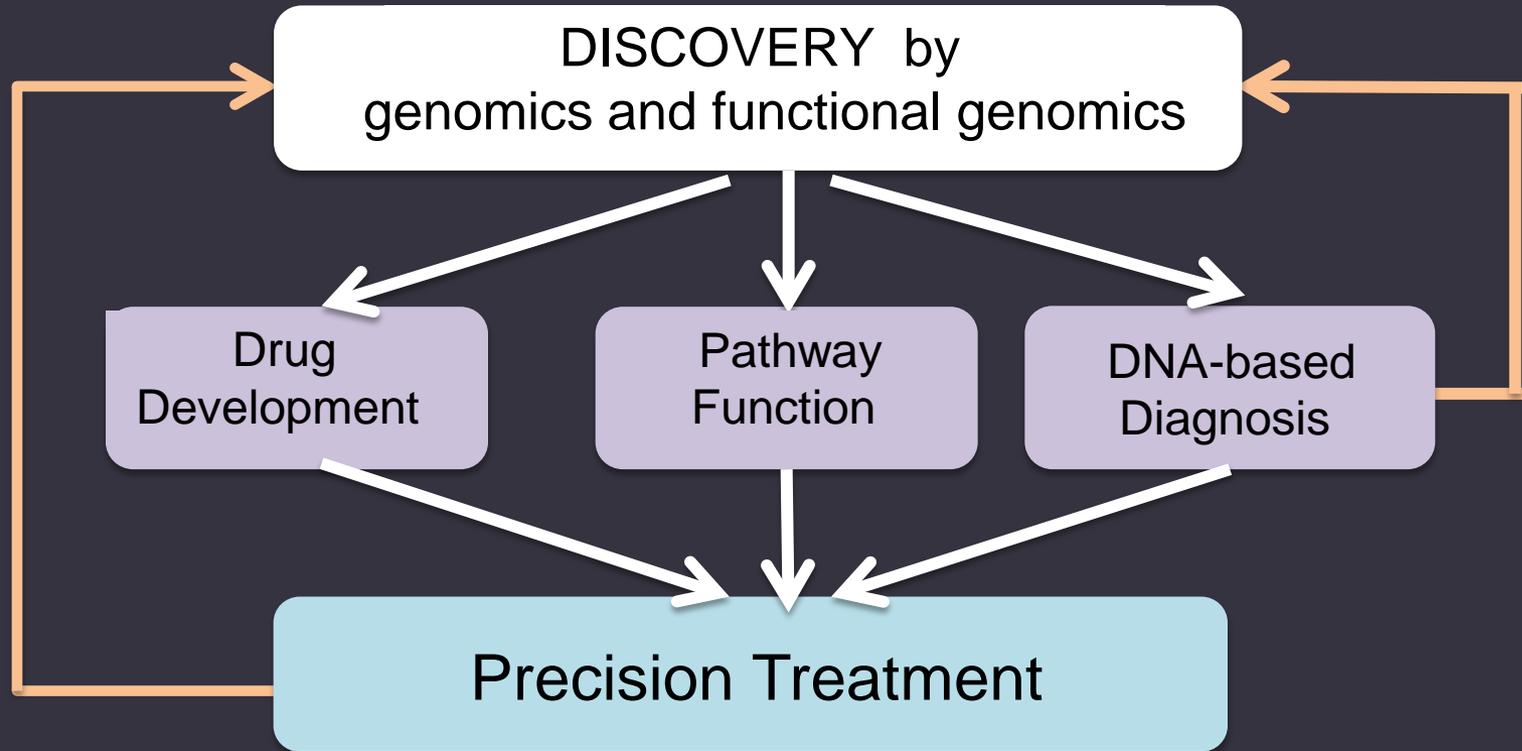


Cancer Genomics 2013-> 2018

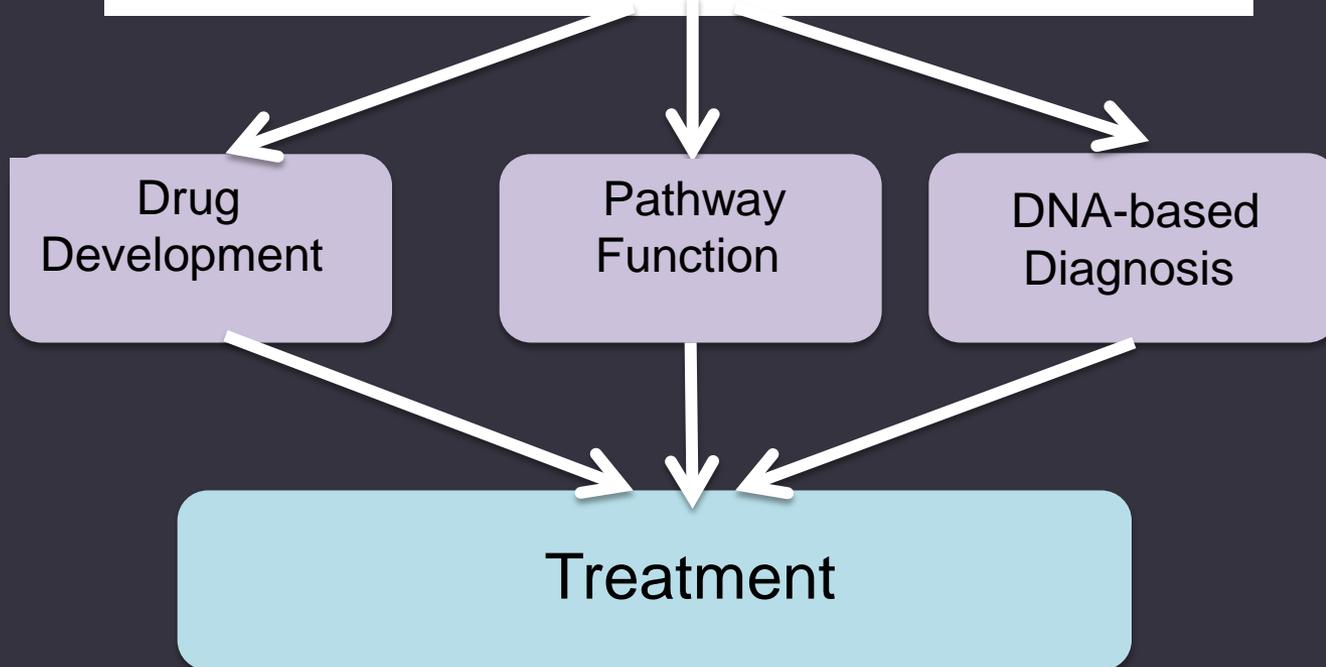


What Bioinformatics & Computational Activities are needed?

What Storage and User Access Resources are needed?

~ Current Genomics 2012-2013

DISCOVERY genomics TCGA/TARGET
20,000 Exomes, RNA, WGS + basic metadata
500GB – 5MGB/case



Current Data storage and ACCESS => CGHub TCGA and TARGET Genomics + some key metadata

- Stores BAM & VCF for TCGA, TARGET and CGAP/CGCI projects
- Access, sorting functions for download
- Designed for 25,000 cases with average of 200 gigabytes per case (some compression anticipated)
- 5 petabytes (5×10^{15}) total, scalable to 20 petabytes
- co-location opportunity/issue



Modified from D. Haussler

Even a “basic” TCGA/TARGET bioinformatics task is NOT simple: Major improvements needed

The spectra of somatic mutations across many tumor types

Mike Lawrence, Gad Getz
Broad Institute of Harvard and MIT

1st Annual TCGA Scientific Symposium
November 17, 2011

Lung cancer

MutSig v0

assuming uniform
bkgd mutation rate
across all genes

$q < 10^{-7}$

- #1 **TP53**
- #2 **KRAS**
- #7 **OR4A15**
- #13 **KEAP1**
- #14 **OR8H2**
- #15 **STK11**
- #17 **OR2T4**
- #25 **OR2T3**
- #31 **OR2T6**
- #48 **CSMD3**
- #49 **OR5D16**
- #55 **RYR2**
- #100 **CSMD1**
- #139 * **PIK3CA**
- #158 **RYR3**
- #159 **MUC16**
- #161 **OR2T33**
- #169 * **NFE2L2**
- #172 **OR10G8**
- #180 **OR2L8**
- #198 **MUC17**
- #217 **TTN**

843 genes

significantly mutated
($q < 0.01$)

* known lung cancer genes
"fishy" genes

improved MutSig

using gene-specific
background mutation rates

- * **STK11** #1
- * **NFE2L2** #4
- * **TP53** #7
- * **KRAS** #8
- * **KEAP1** #11
- * **PIK3CA** #12

$q < 10^{-5}$

52 genes

significantly mutated

($q < 0.01$)

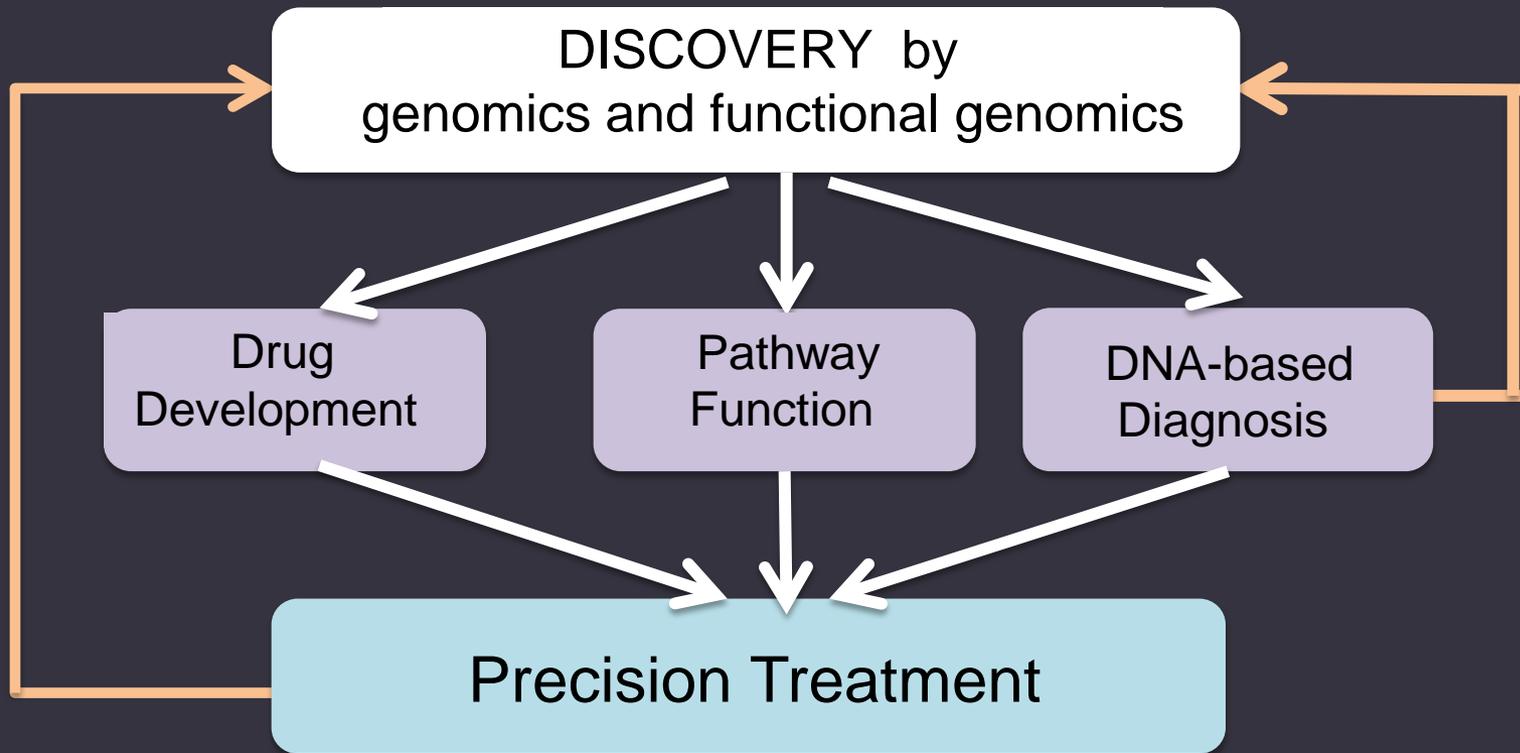
- * **OR8H2** #181
- OR5T2** #276
- OR10J3** #334
- CSMD3** #388
- MUC17** #2614
- RYR2** #2898
- CSMD1** #4482
- TTN** #4825
- MUC16** #5650
- RYR3** #11496

$q \sim 0.2$

$q = 1$

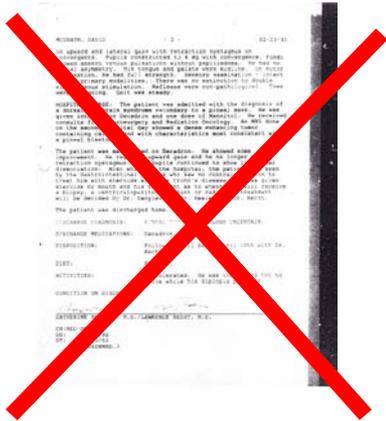
* most significant
olfactory receptor

Cancer Genomics 2013-> 2018



What is needed to make the orange arrows real?

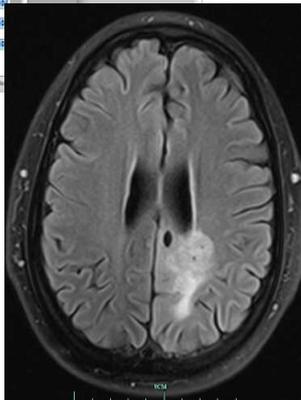
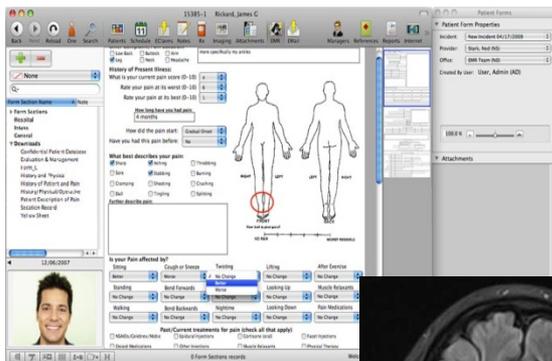
Future: Genotype-Phenotype Correlations Compute via Clinical Data Records and Genomics



- Scanned medical records are NOT computable EMRs

- True EMRs (e.g. Vanderbilt, VA, etc) years of data for 100s of standard elements

Average patient: 1-100 MB work of clinical data (text/data elements)

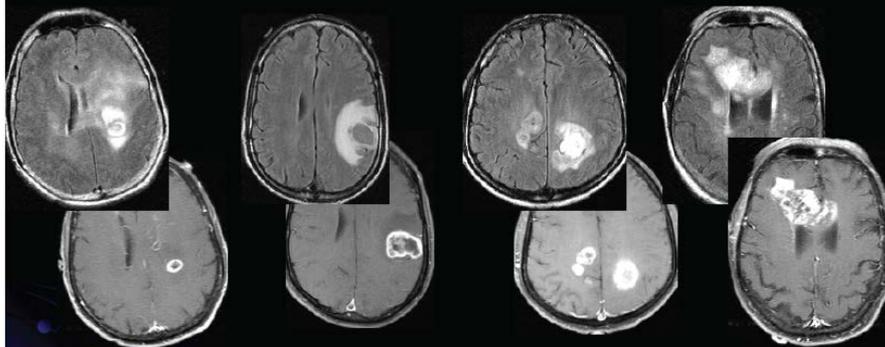


- Images, other data types, expand order of magnitude to ~1-2TB

Expression-based Classes Reveal Different MRI Features



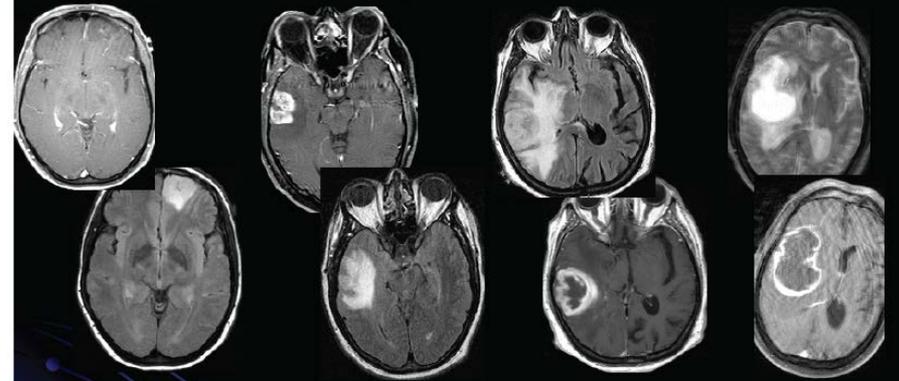
f5 – Proportion Enhancing



(3) < 5% (4) 6-33% (5) 34-67% (6) 68-95%

Visually, when scanning through the entire tumor volume, what proportion of the entire tumor would you estimate is enhancing. (Assuming that the entire abnormality may be comprised of: (1) an enhancing component, (2) a non-enhancing component, (3) a necrotic component and (4) a edema component.)

f7 – Proportion Necrosis



(2) None (3) < 5% (4) 6-33% (5) 34-67%

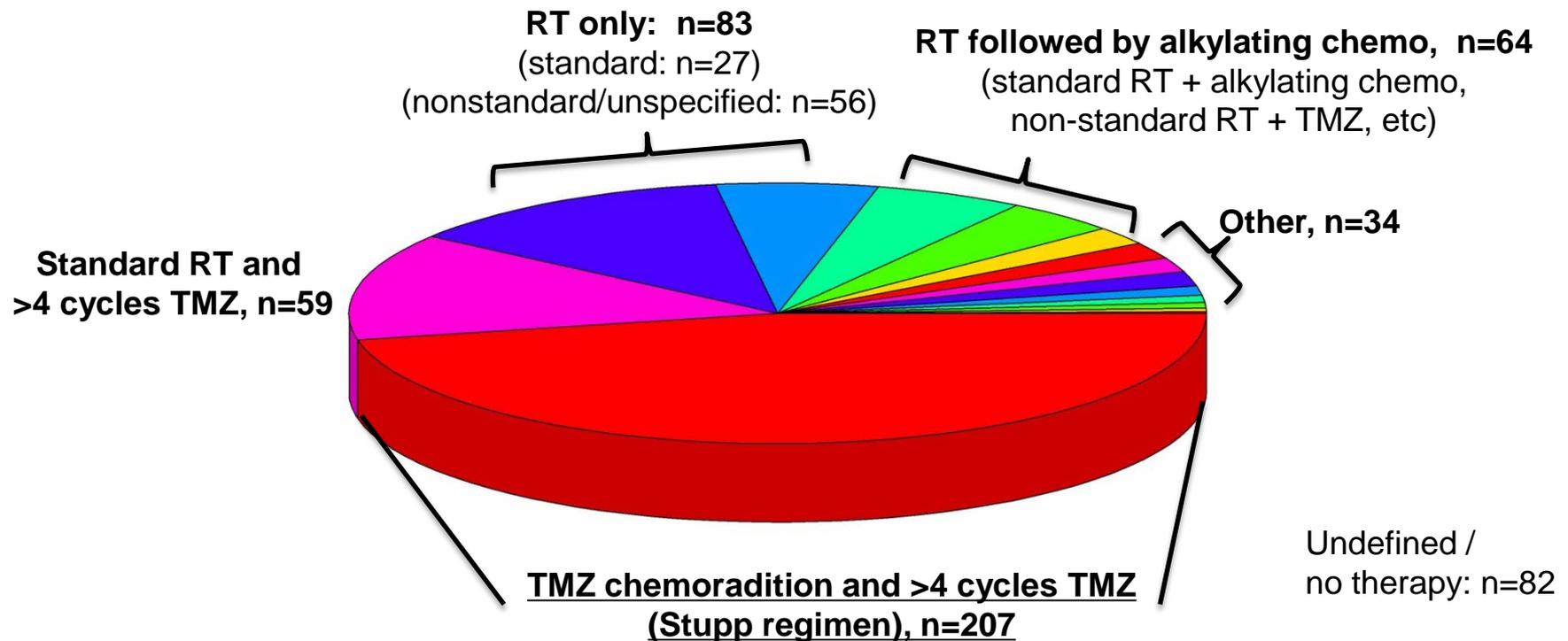
Visually, when scanning through the entire tumor volume, what proportion of the tumor is estimated to represent necrosis. Necrosis is defined as a region within the tumor that does not enhance or shows markedly diminished enhancement, is high on T2W and proton density images, is low on T1W images, and has an irregular border. (Assuming that the entire abnormality may be comprised of: (1) an enhancing component, (2) a non-enhancing component, (3) a necrotic component and (4) a edema component.)

- *TP53* mutant tumors: smaller mean tumor sizes ($p=0.002$) T2-weighted or FLAIR images.
- *EGFR* mutant tumors: significantly larger than *TP53* mutant tumors ($p=0.0005$).
- High level *EGFR* amplification: associated with >5% enhancement and >5% proportion of necrosis ($p < 0.01$).

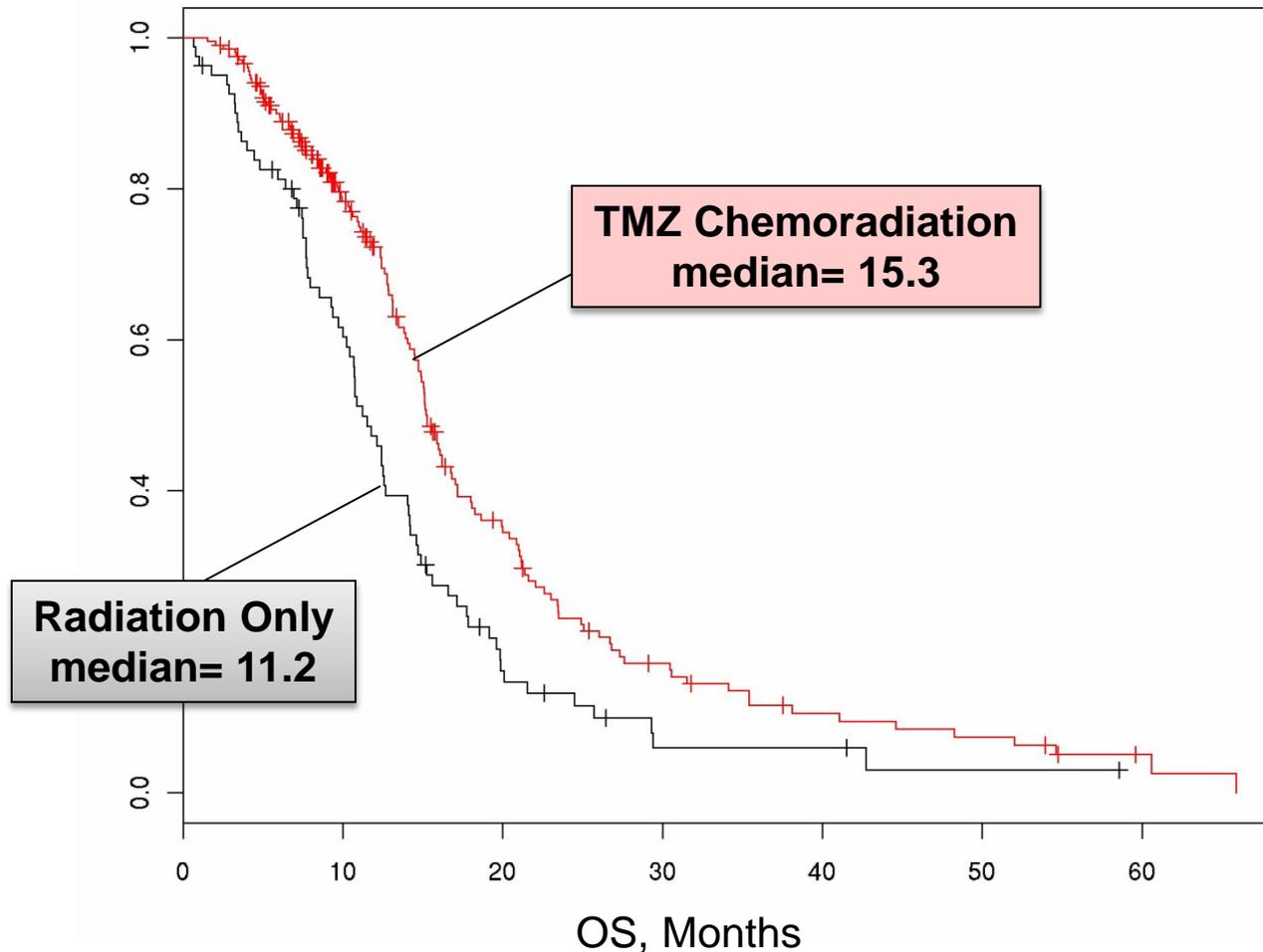
Importance of Clinical Data



- Case study: GBM treatment data in TCGA
- 207 cases with TMZ chemoradiation



Clinical Data -> toward genomic correlates of different treatment outcomes



Future Tumorbases: When? How different from CG-Hub?

Sequence Data
Previous Driver to Big –
Shrink?

EMR Data – more and
bigger!

Alternate structures –
Secure Clouds?

