Mutation Discovery in TSP and TCGA

- Lung Adenocarcinoma and Glioblastoma

Li Ding

Medical Genomics Group

Washington University Genome Sequencing Center

lding@watson.wustl.edu

Advance in the Discovery of Somatic Mutations in Cancer

1. Known somatic mutations in COSMIC (Nov 2007 stats)

-Cancer samples 246,369 (Cancer types 69)
-Somatic mutations 51,599 (Unique 9,559)
-Genes with mutation found 4,762

2. Recent advance of large-scale cancer mutation studies

-Screen large number of genes in limited cancer samples
-Screen a selected subset of genes in different cancer types
Greenman *et al.* Nature 2007; Sjöblom *et al.* Science, 2006

3. TSP and TCGA projects aim at further advancing mutational discovery in cancers

-Larger number of cancer types
-Larger sample size for each cancer type
-Integration of different platforms (SNP array, Mutation discovery, expression profiling)
-More comprehensive analysis

➢ TSP Lung Adenocarcinoma Study

❖188 tumor samples and 630 genes have been through PCR-based re-sequencing. 384 lung adenocarcinoma tumor samples and matched normals were used for Affy SNP array study. RNAs from 76 tumor samples have been used for Affy gene expression array.

❖Moving into TSP part B: large scale re-sequencing using NexGen sequencers.

➢ TCGA Glioblastoma Study

❖100 glioblastoma samples/matched normals and 605 phase I genes have been through PCR-based re-sequencing. The same set of samples have been analyzed by other platforms such as SNP array, gene expression array, and so on.

❖Moving into re-sequencing with phase II genes and regions (~700).

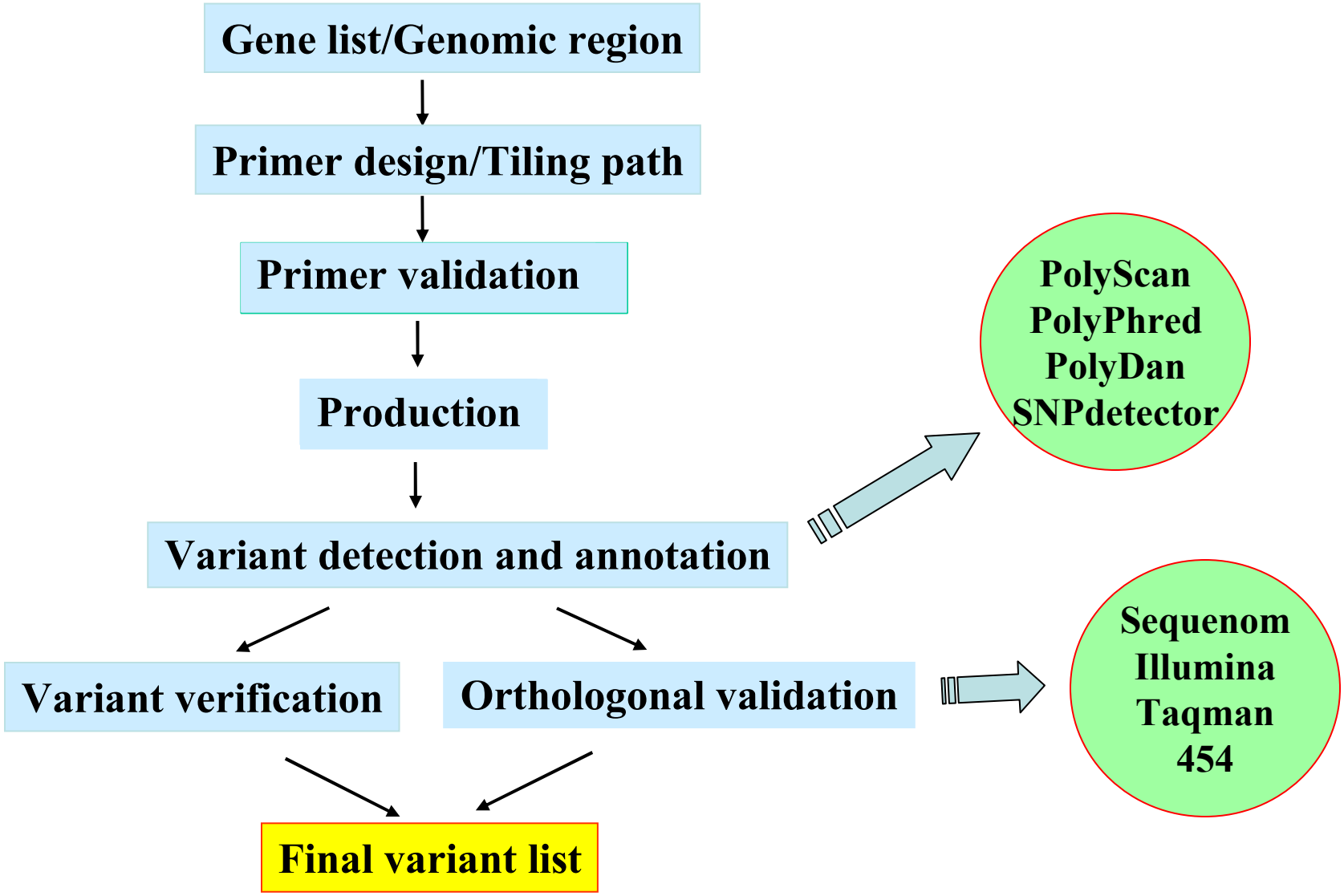
Outline

- **Pipeline**
  - Overview of medical sequencing pipeline
  - Overview of medical sequencing analysis pipeline

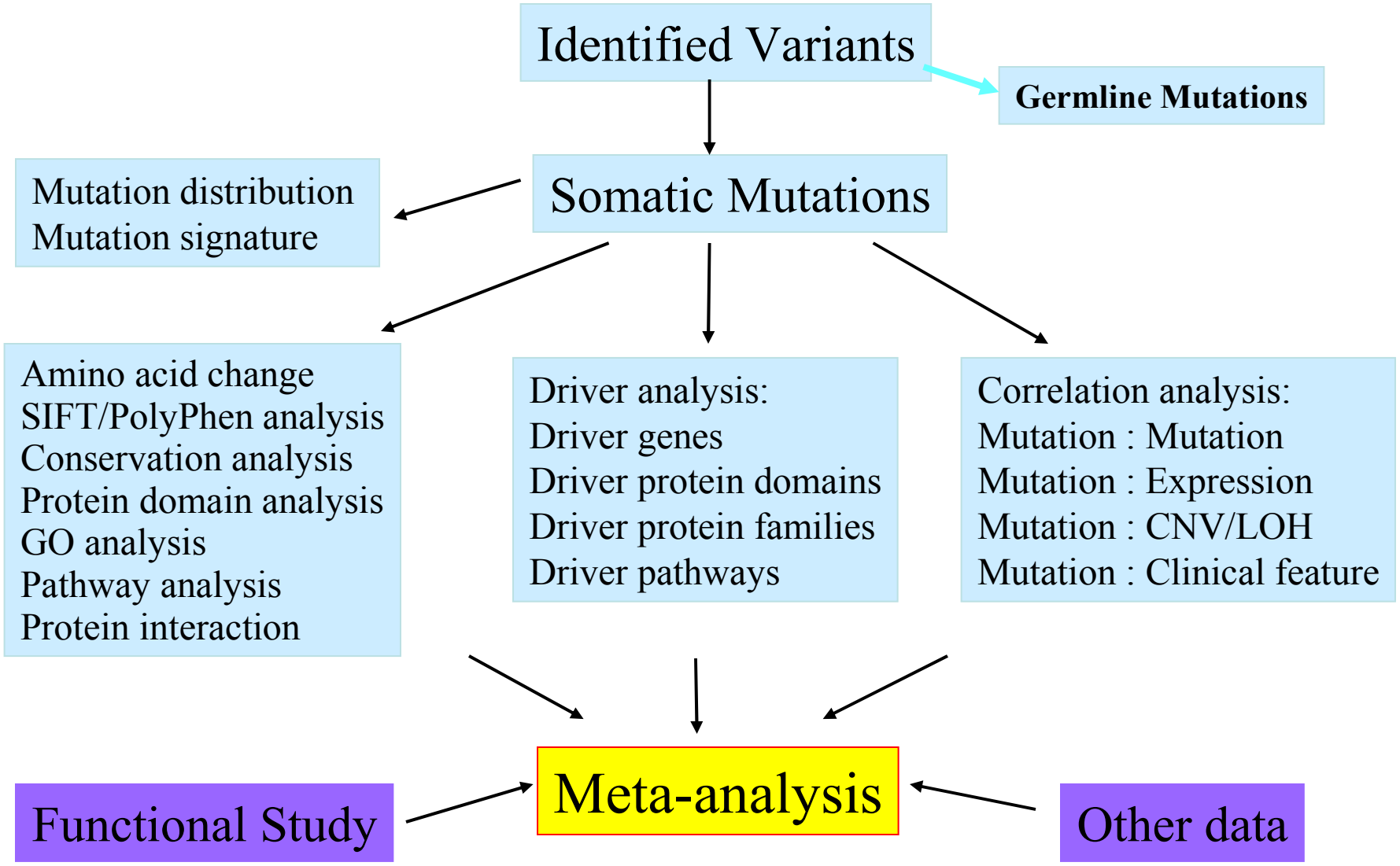
- **Analysis of lung adencarcinoma**
  - Mutation stats, distribution, and signature
  - Driver vs. passenger analysis
  - Correlation among mutations
  - Correlation between mutations and LOH/CNV
  - Correlation between mutations and clinical features

- **Analysis of glioblastoma and comparison to lung adenocarcinoma**
  - Mutation stats and cross-center comparison
  - Mutation distribution and signature
  - Comparison between two cancer types

PCR-based Medical Sequencing Pipeline



Medical Sequencing Mutation Analysis Pipeline



TSP Lung Adenocarcinoma Project

-Verified or Validated Somatic Mutations Identified in Lung Adenocarcinoma

Shared 61 genes: 316 Point Mutations and 21 Indels

WUGSC 190 genes: 256 Point Mutations and 13 Indels

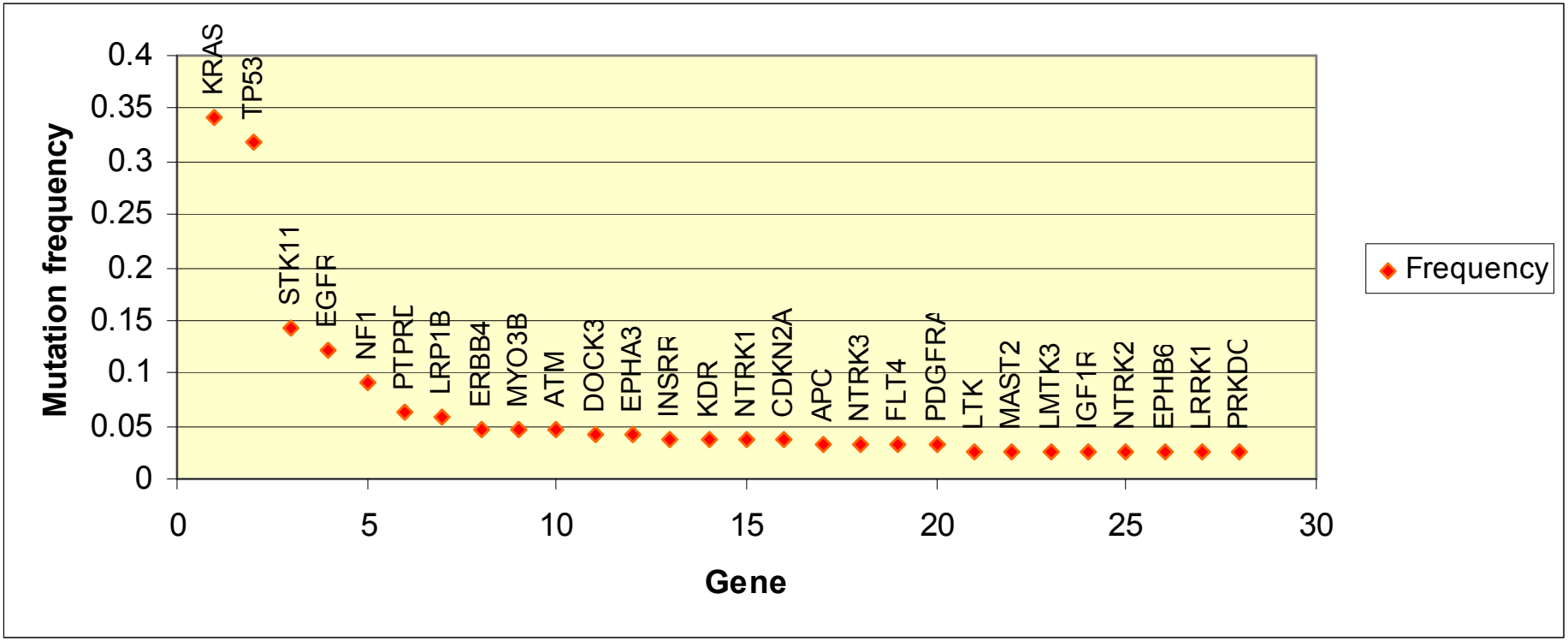
Broad 194 genes: 118 Point Mutations

Baylor 185 genes: 210 Point Mutations

Total 630 genes: **900 Point Mutations and 34 Indels**

Mutation Frequencies in Lung Adenocarcinoma

--KRAS and TP53 Are Mutated in About 1/3 of Tumor Samples

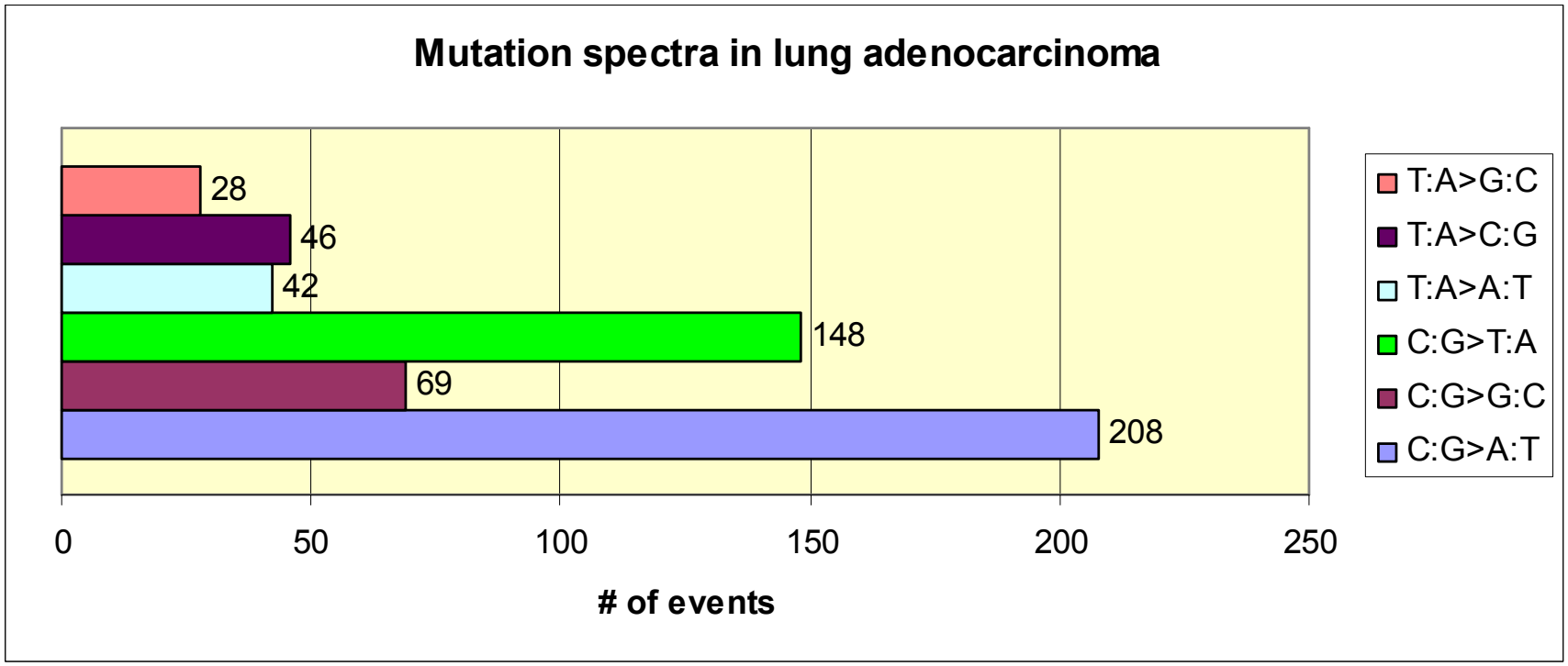


- Indels have not been included in the analysis

- WU/Broad/Shared genes used for the analysis

Mutation Signatures in Lung Adenocarcinoma

-Mutation Signature Potentially Determined by the Specific Carcinogen in Smoke

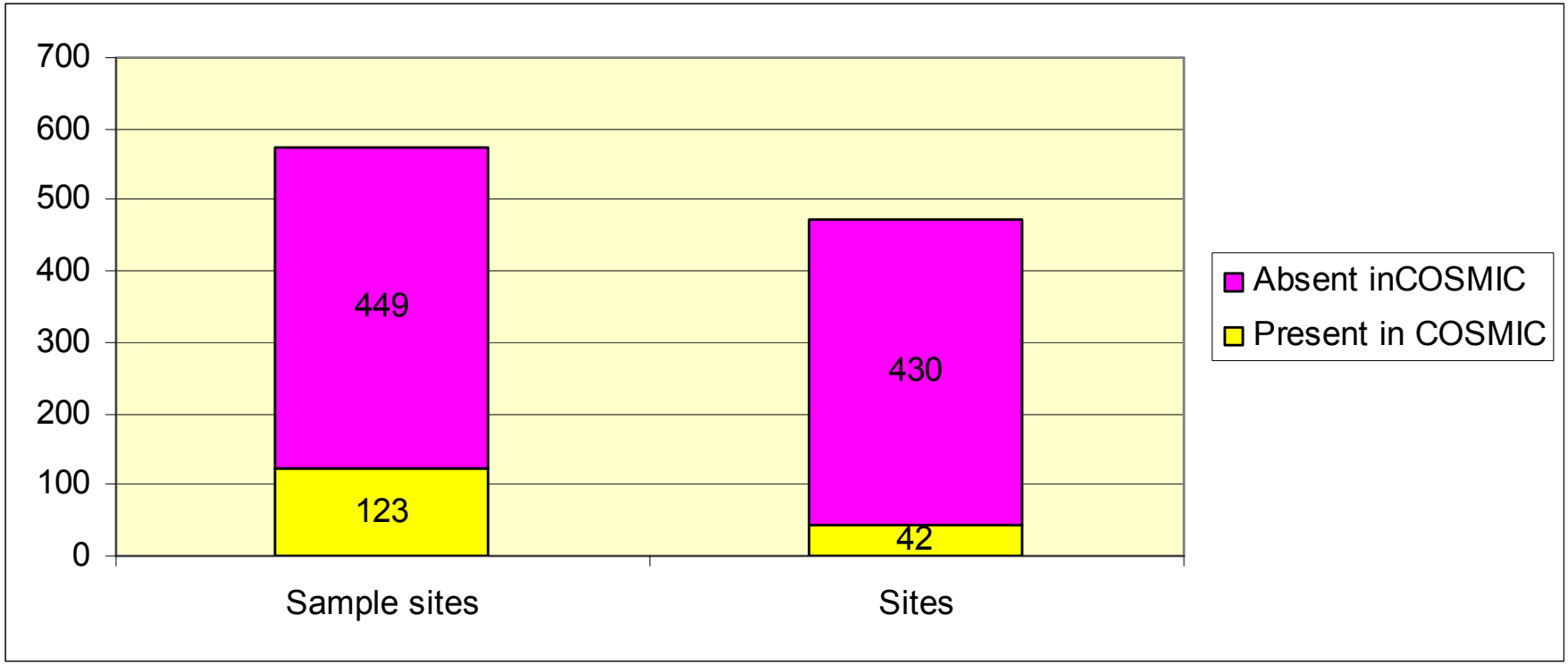


- C:G>T:A transition and C:G>A:T transversion are most frequent

❖WU/Shared genes used for the analysis

90% Mutant Sites Identified in Lung Adencarcinoma Are Novel

-Known and Novel Mutations Found in TSP Samples



❖Indels not included in this graph

❖WU/Shared genes used for the analysis

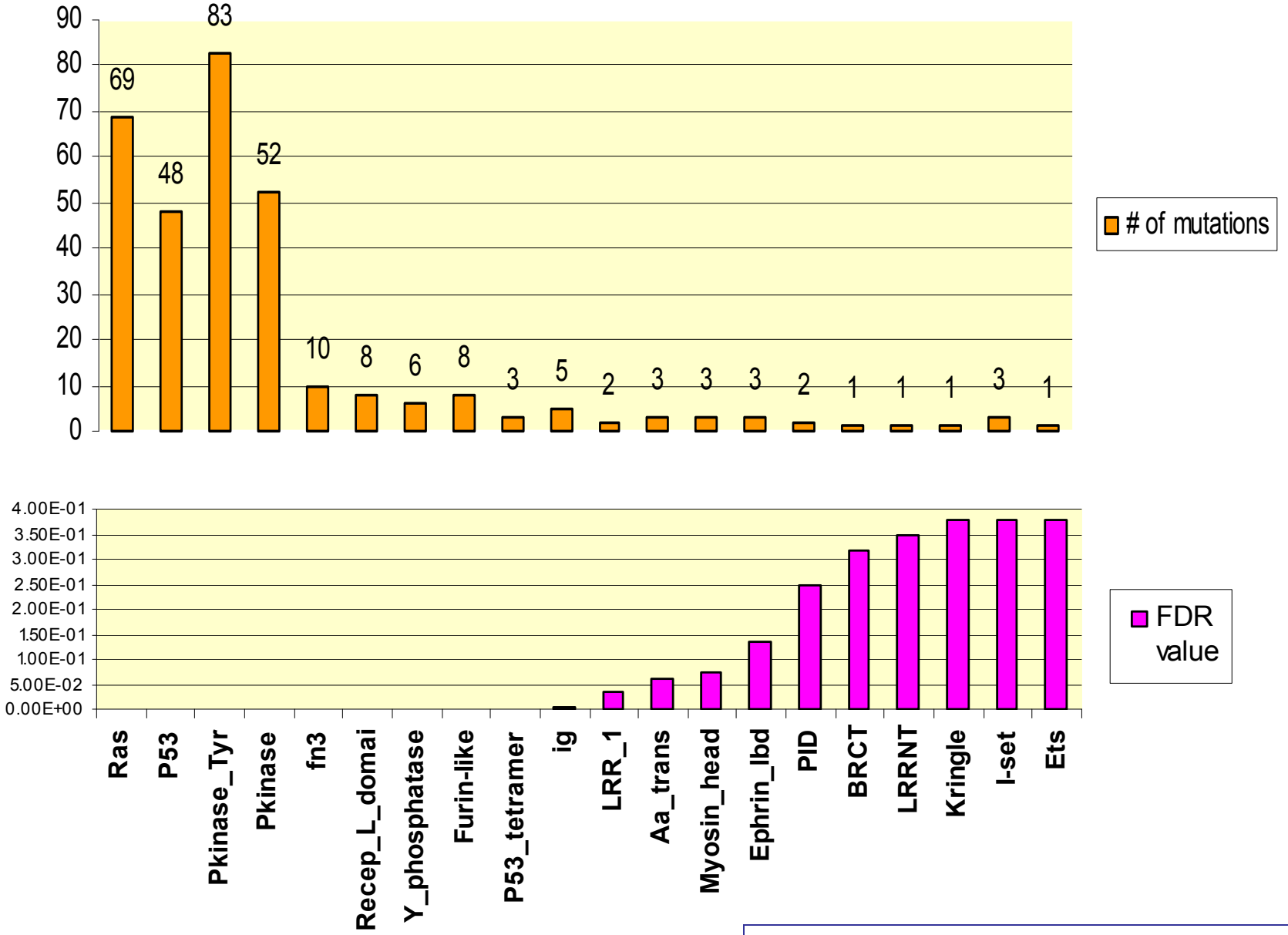
Identify "Drivers" in Lung Cancer

Gene	# of mutations	# of targeted bases	P_value	FDR	Bonferroni
KRAS	64	144196	1.61E-124	3.55E-122	3.55E-122
TP53	60	266020	2.55E-099	2.82E-097	5.63E-097
STK11	27	281624	9.90E-036	7.29E-034	2.19E-033
EGFR	23	840172	1.18E-018	6.54E-017	2.62E-016
CDKN2A	7	197024	2.08E-007	9.17E-006	4.59E-005
NF1	17	1816080	3.10E-007	1.14E-005	6.84E-005
PTPRD	12	1230272	1.08E-005	0	0
EPHA3	8	518725	1.33E-005	0	0
ERBB4	9	852016	7.30E-005	0	0.02
NTRK1	7	540124	0	0	0.03
YO3B	9	926276	0	0	0.03
INSRR	7	811032	0	0.03	0.32
NTRK3	6	603668	0	0.03	0.34
PTEN	4	257936	0	0.03	0.43
KDR	7	874388	0	0.03	0.48
PDGFRA	6	693720	0	0.04	0.68
ZMYND10	4	307192	0	0.05	0.81
SLC38A3	4	337272	0.01	0.06	1
ATM	9	1590725	0.01	0.06	1
NTRK2	5	557232	0.01	0.06	1
LTK	5	579792	0.01	0.07	1
DOCK3	8	1358112	0.01	0.07	1
EPHB6	5	624348	0.01	0.09	1
TFDP1	3	216767	0.01	0.09	1
LRP1B	11	2425277	0.01	0.1	1
FLT4	6	922516	0.01	0.1	1

- Estimated background mutation rate (BMR) 2.00e-6
- False Discovery Rate (FDR) < 0.1

- WU/Broad/Shared genes used for the analysis

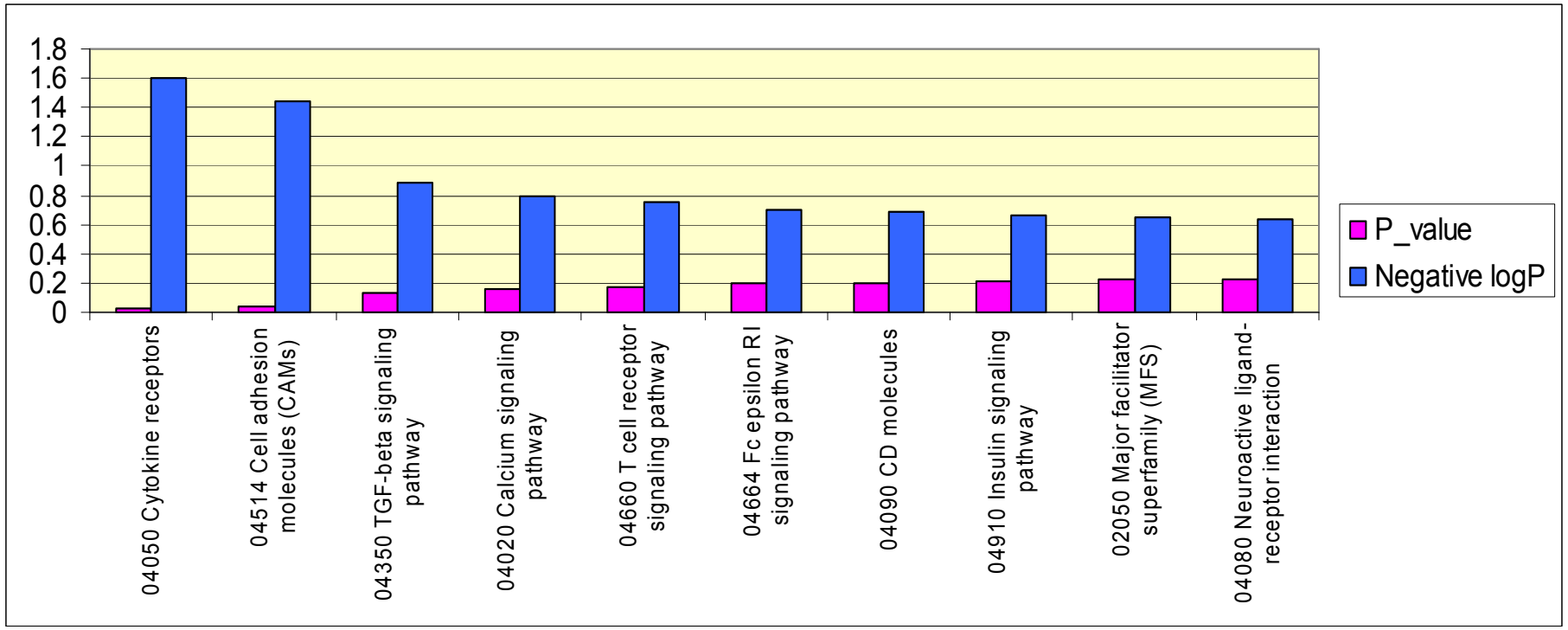
Driver Protein Domains Involved in Lung Adenocarcinoma



❖WU/Broad/Shared genes used for the analysis

Top 10 KEGG Level 3 Pathways in Lung Adenocarcinoma

-Cytokine Receptors and Cell Adhesion Are the Top 2 Pathways Mutated in Lung Adenocarcinoma



❖WU/Broad/Shared genes used for the analysis

Gene 1	Gene 2	CC	P_value	FDR	Bonferroni
FYN	PRKDC	0.48903	2.65E-11	9.32E-09	3.73E-08
ERBB2	LRRK1	0.458802	5.73E-10	8.06E-08	8.06E-07
FGFR4	PDGFRA	0.458802	5.73E-10	8.06E-08	8.06E-07
NTRK2	PDGFRA	0.458802	5.73E-10	8.06E-08	8.06E-07
EPHA3	LTK	0.453912	9.17E-10	1.07E-07	1.29E-06
FYN	LMTK3	0.437188	4.32E-09	4.34E-07	6.08E-06
INSRR	PDGFRA	0.428862	9.08E-09	7.98E-07	1.28E-05
ERBB2	NTRK1	0.414721	3.06E-08	2.39E-06	4.30E-05
LMTK3	PDGFRA	0.410403	4.38E-08	3.08E-06	6.16E-05
NF1	PDGFRA	0.408854	4.98E-08	3.18E-06	7.00E-05
FGFR4	NTRK2	0.38125	4.37E-07	2.56E-05	0.000614
LRP1B	PRKDC	0.377964	5.58E-07	3.02E-05	0.000785
PDGFRA	PTPRD	0.365382	1.40E-06	7.01E-05	0.001962
LRRK1	NTRK1	0.343401	6.31E-06	0.000296	0.008876
DOCK3	LTK	0.330457	1.46E-05	0.000641	0.020499
DOCK3	FLT4	0.316538	3.44E-05	0.00121	0.048383
LMTK3	TP53	0.314988	3.78E-05	0.001264	0.053102
CDKN2A	ERBB2	0.308348	5.59E-05	0.001788	0.078651
EGFR	KRAS	-0.30031	8.89E-05	0.002505	0.125004
FGFR4	NF1	0.300273	8.91E-05	0.002505	0.12524
NF1	NTRK2	0.300273	8.91E-05	0.002505	0.12524
LRP1B	TP53	0.296045	0.000113	0.002747	0.158928
ERBB2	MAST2	0.296001	0.000113	0.002747	0.159313
ERBB2	PRKDC	0.296001	0.000113	0.002747	0.159313
PDGFRA	PRKDC	0.296001	0.000113	0.002747	0.159313
LRP1B	PDGFRA	0.290883	0.00015	0.003267	0.211516
EPHA3	EPHB6	0.289306	0.000164	0.003267	0.230561

Concurrence and Mutual Exclusion among the Mutations -Correlations between Mutations

- PRKDC, ERBB2, and PDGFRA show strong positive correlation with other genes

- EGFR and KRAS display strong negative correlation

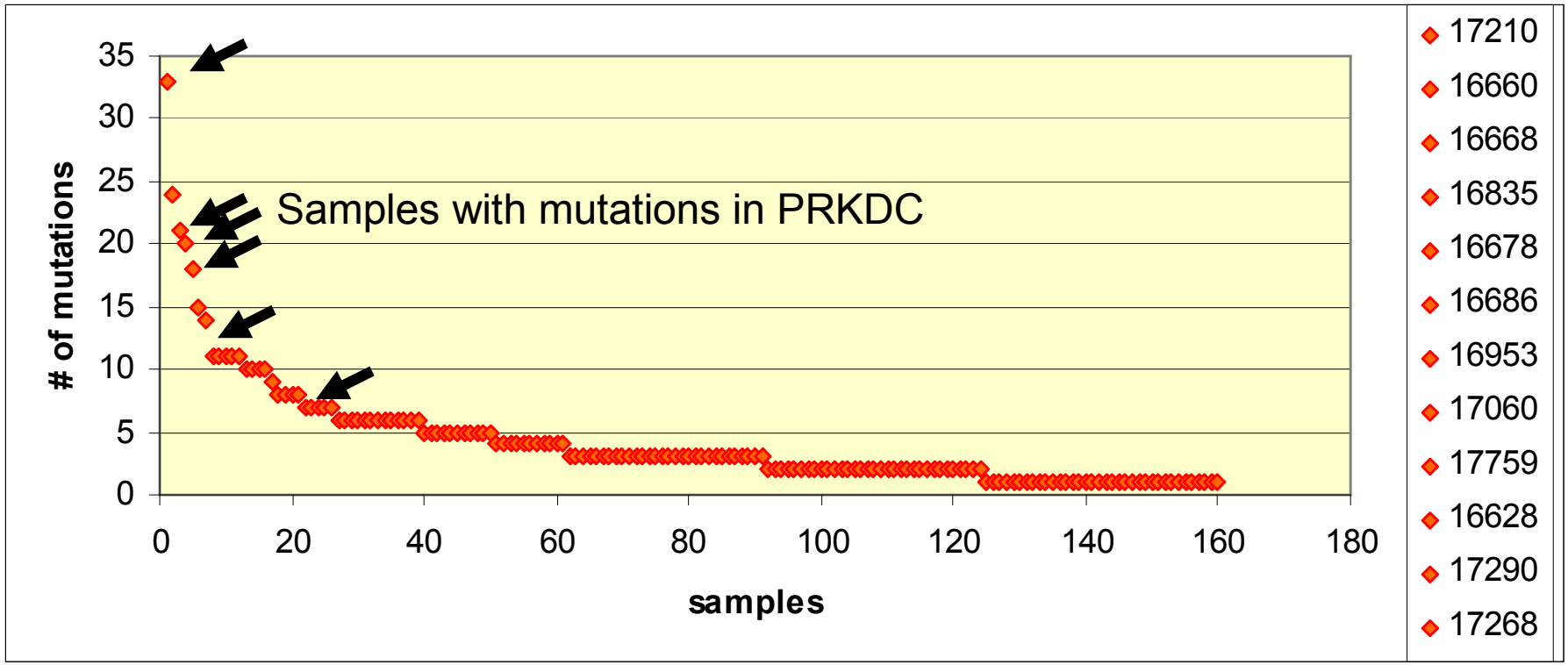
*Only genes with >=5 mutations used for this analysis

*CC: correlation coefficients

❖WU/Broad/Shared genes used for the analysis

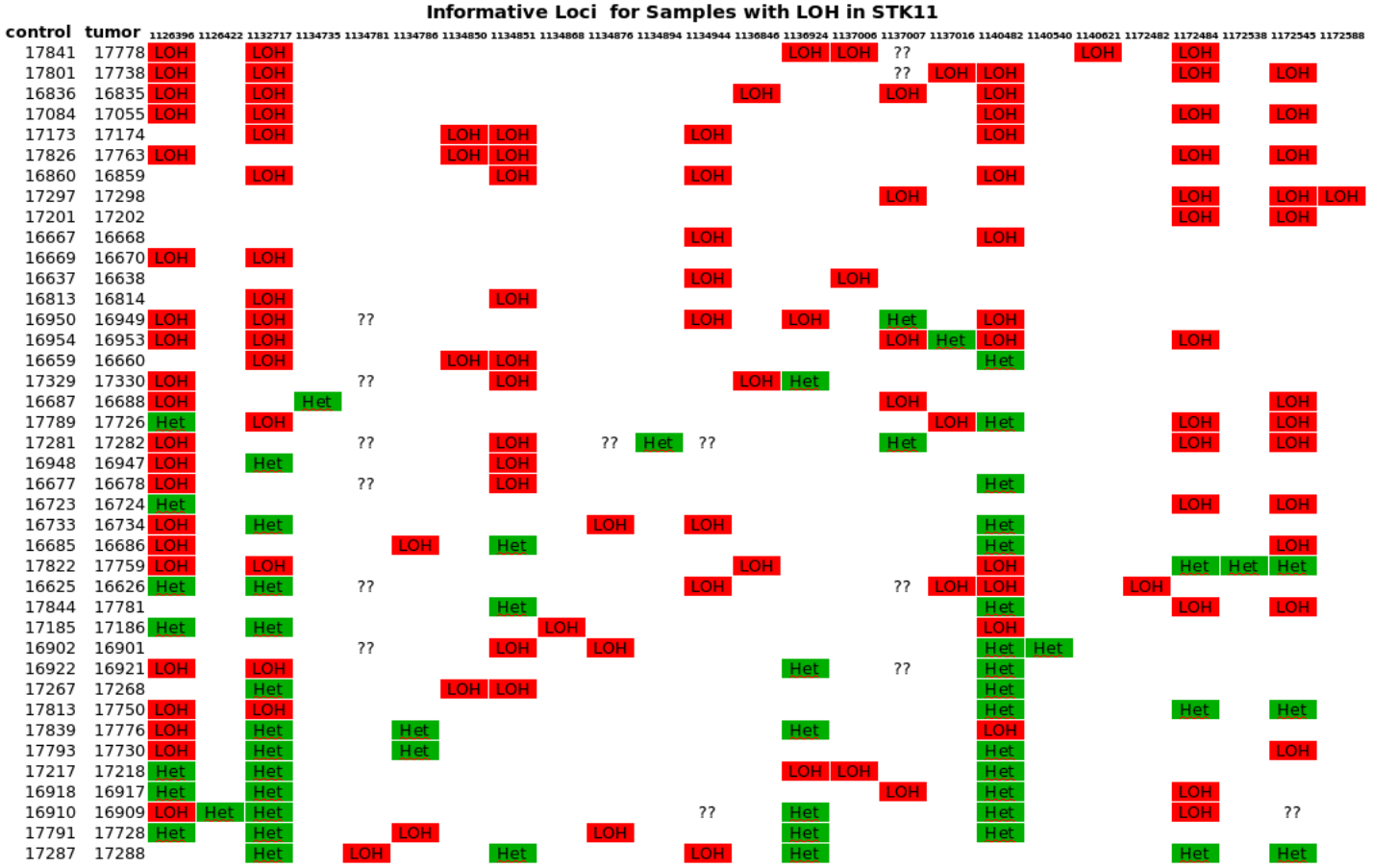
Highly Mutated Samples Have Mutations in PRKDC

-Distribution of Somatic Mutations in Lung Adenocarcinoma Samples



- PRKDC encodes the catalytic subunit of a nuclear DNA-dependent serine/threonine protein kinase (DNA-PK).
- It is involved in the repair of double-strand breaks in DNA in which the two broken ends are rejoined with little or no sequence complementarity.

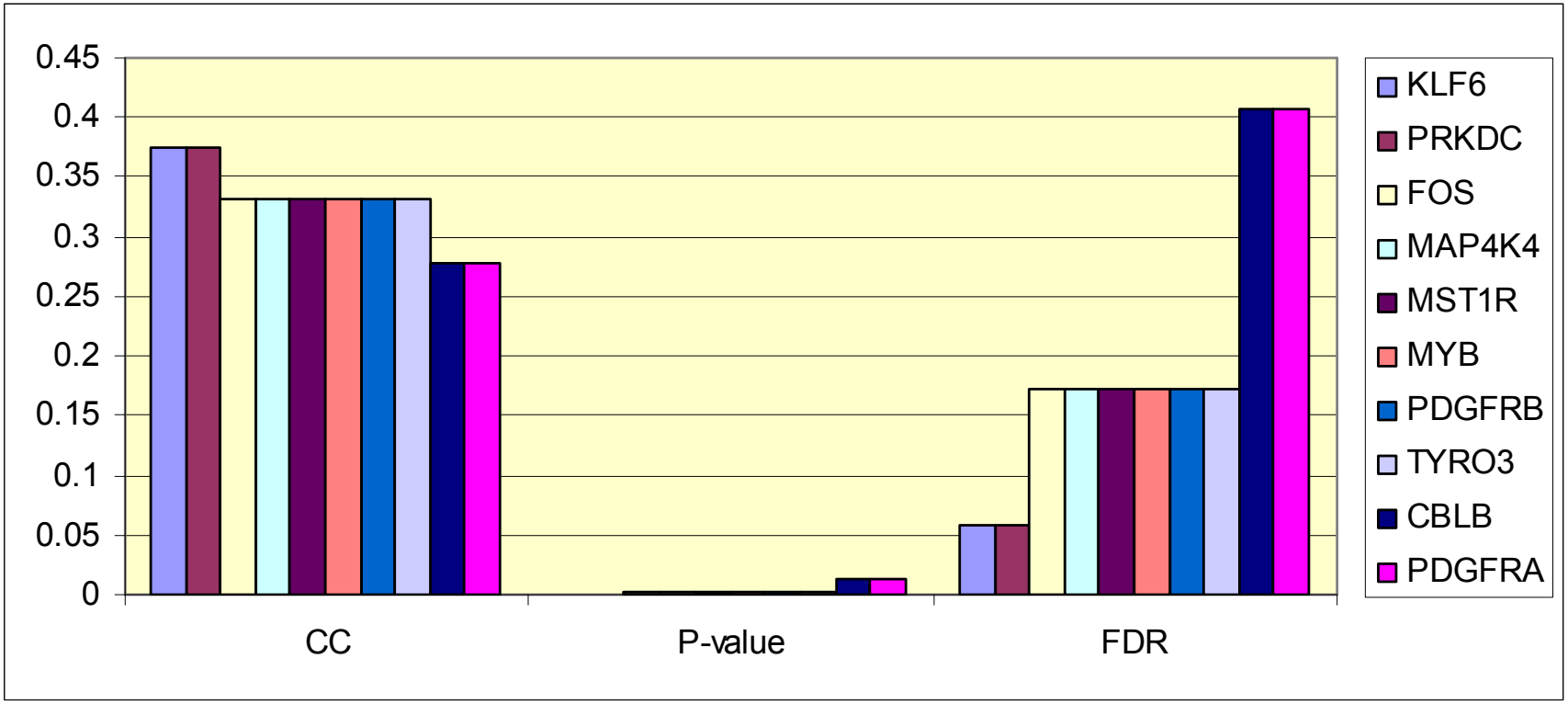
Driver Gene STK11 Is Frequently Targeted by Both Point Mutations and LOH/CNV



- STK11 is a serine/threonine kinase. Mutations in STK11 are associated the growth of polyps in Peutz-Jeghers syndrome.
- Identified **27** nonsynonymous SNPs and **4** Indels in STK11 in 188 lung adenocarcinoma samples.
- Third highest mutated gene among the genes screened so far in lung adencarcinoma.
- 25 Polymorphic STK11 SNP sites sequenced to identify samples with LOH.

Associations between Mutations and Clinical Features

- Mutations in KLF6 and PRKDC Are Positively Correlated with Tumor Stage



*CC: correlation Coefficients

❖WU/Shared genes used for the analysis

Associations between Mutations and Clinical Features

- Mutations in TP53 show significant positive correlation with tumor grade.

- Mutations in PRKDC, TP53, STK38L, and TSC2 (tuberous sclerosis 2) correlate with the solid tumor histological subtype of lung adenocarcinoma.

- High correlation of mutations in EGFR, MST4, and KSR1 with never smoker and mutations in KRAS with smokers.

❖WU/Shared genes used for the analysis

Summary of TSP Study

- Completed the main production and mutation screening for 630 genes in 188 lung adenocarcinoma samples

- Identified driver genes, domains, families, and pathways involved in lung tumorogenesis

- Revealed the genes with concurrent or exclusive mutation patterns

- Driver genes may harbor both point mutations and LOH/CNV

- Identified correlations between clinical features and genetic mutations

- Prioritized key mutations for functional study

TCGA Glioblastoma Project
-What Have We Learned from TSP?

- **Important to sequence tumors and matched normals**
  - Determine somatic status based on sequencing data alone
  - Discover LOH events based on sequencing data
  - Identify low level mutations by comparing tumor and normal trace data
  - Identify somatic silent mutations to estimate background mutation rate

- **Important to perform gap filling to gain high percentage of quality coverage cross the target regions**

TCGA Production Progress

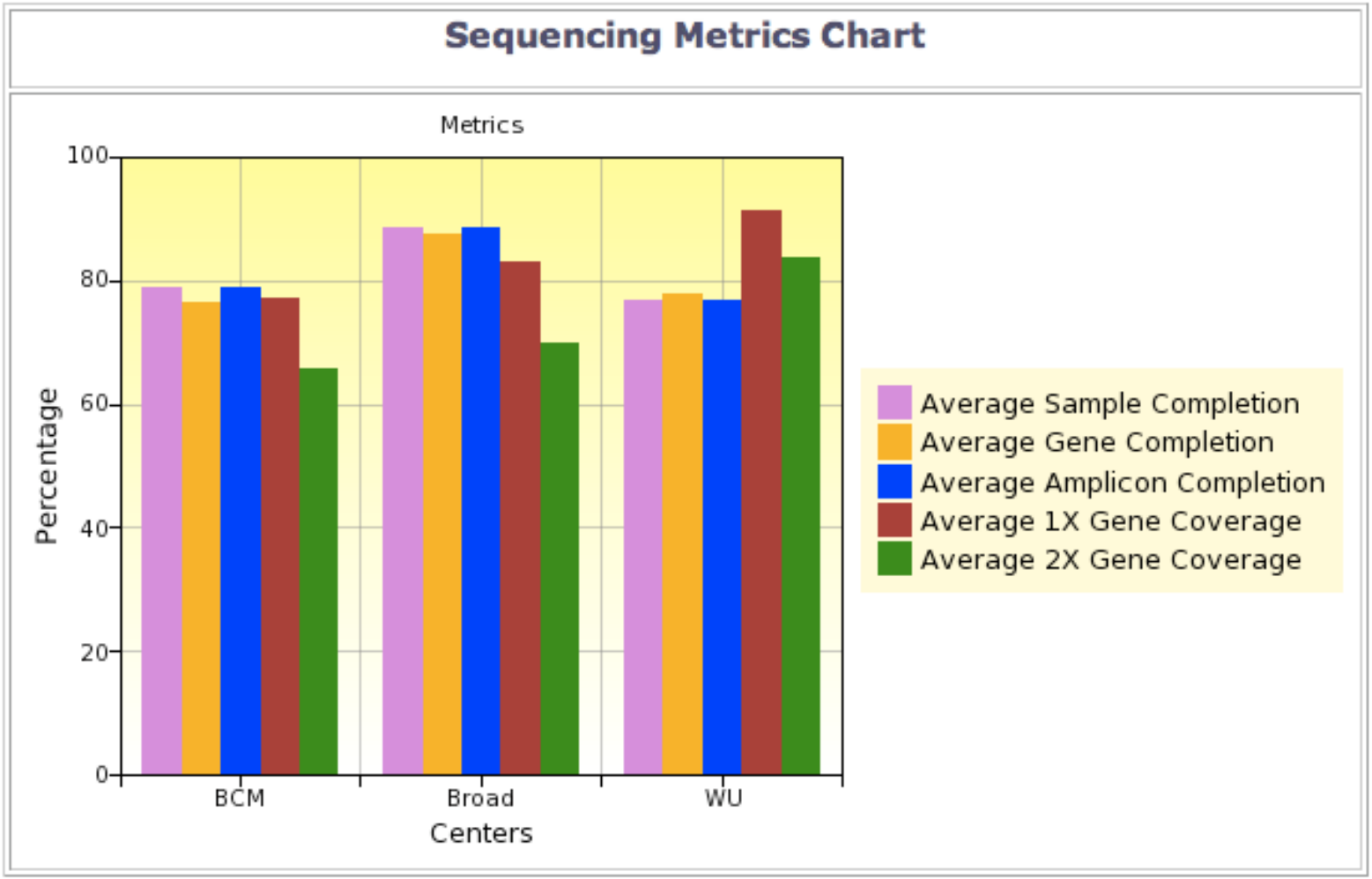
-Global Re-sequencing Production Statistics

Global Completion Statistics Table				
	BCM	Broad	WU	Total Unique
1. Samples targeted by NCI	500	500	500	500
2. Samples delivered to centers	223	223	223	223
3. Minimum number of Samples to be sequenced *	193	193	193	193
4. Samples reported	188	134	202	210
5. Genes Targeted	230	217	201	605
6. Estimated number of ROIs using Refseq Genes	2,858	2,826	2,824	8,028
7. Reported ROIs	2,836	2,815	2,904	8,555
8. Reported Amplicons	3,121	3,055	3,831	10,007
9. Projected ROI-Samples	547,348	543,295	586,608	1,716,511
10. Completed ROIs	456,670	369,354	586,406	1,412,430

***Minimum number of Samples to be sequenced. 96 unique tumors, 96 normals and 1 control.**

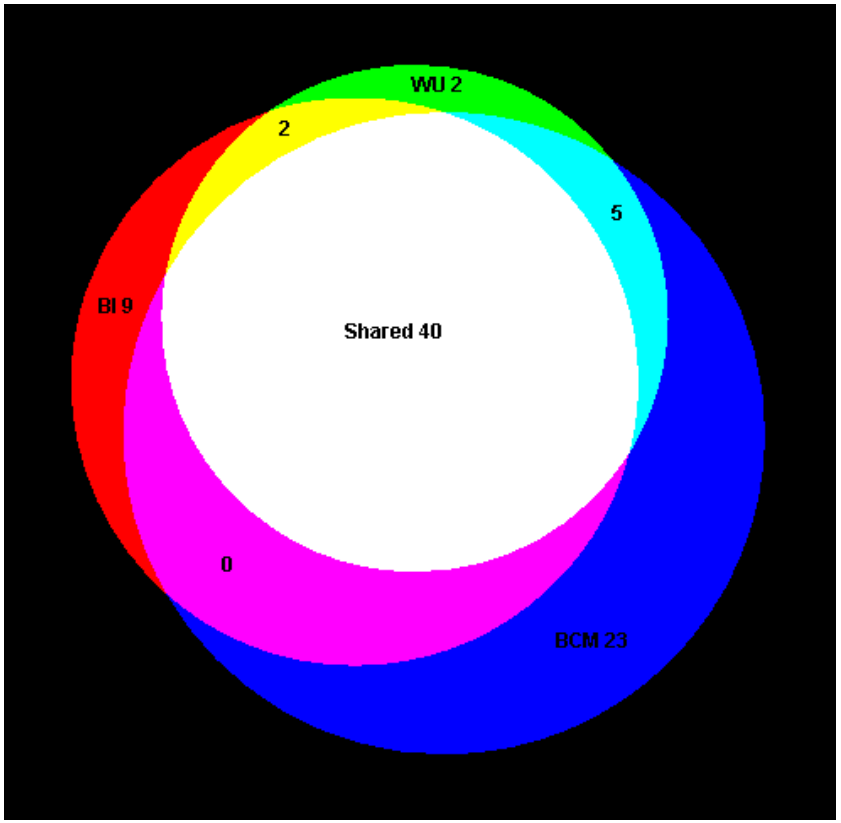
From Genboree

An Average of over 80% 1X Gene Coverage Has Been Achieved by Three Centers



From Genboree

Comparison of Candidate Somatic Mutations Detected by Three Centers in Shared 20 Genes



Sample-site-based comparison

TCGA Batch 1-3 samples:

WU total: 49
Broad total: 51
BCM total: 68
WU_Broad_in_common: 42
WU_BCM_in_common: 45
Broad_BCM_in_common: 40
WU_unique: 2
Broad_unique: 9
BCM_unique: 23*
All_three_centers_in_common: 40

***4 Baylor unique might be real**

Sample/site comparison requires that both the variant site and the variant sample to be identical to be considered as an overlap.

TCGA Glioblastoma

Candidate Somatic Mutations Identified in Glioblastoma

Shared 20 genes: 81 point mutations and 6 Indels

WUGSC 179 genes: 281 point mutations

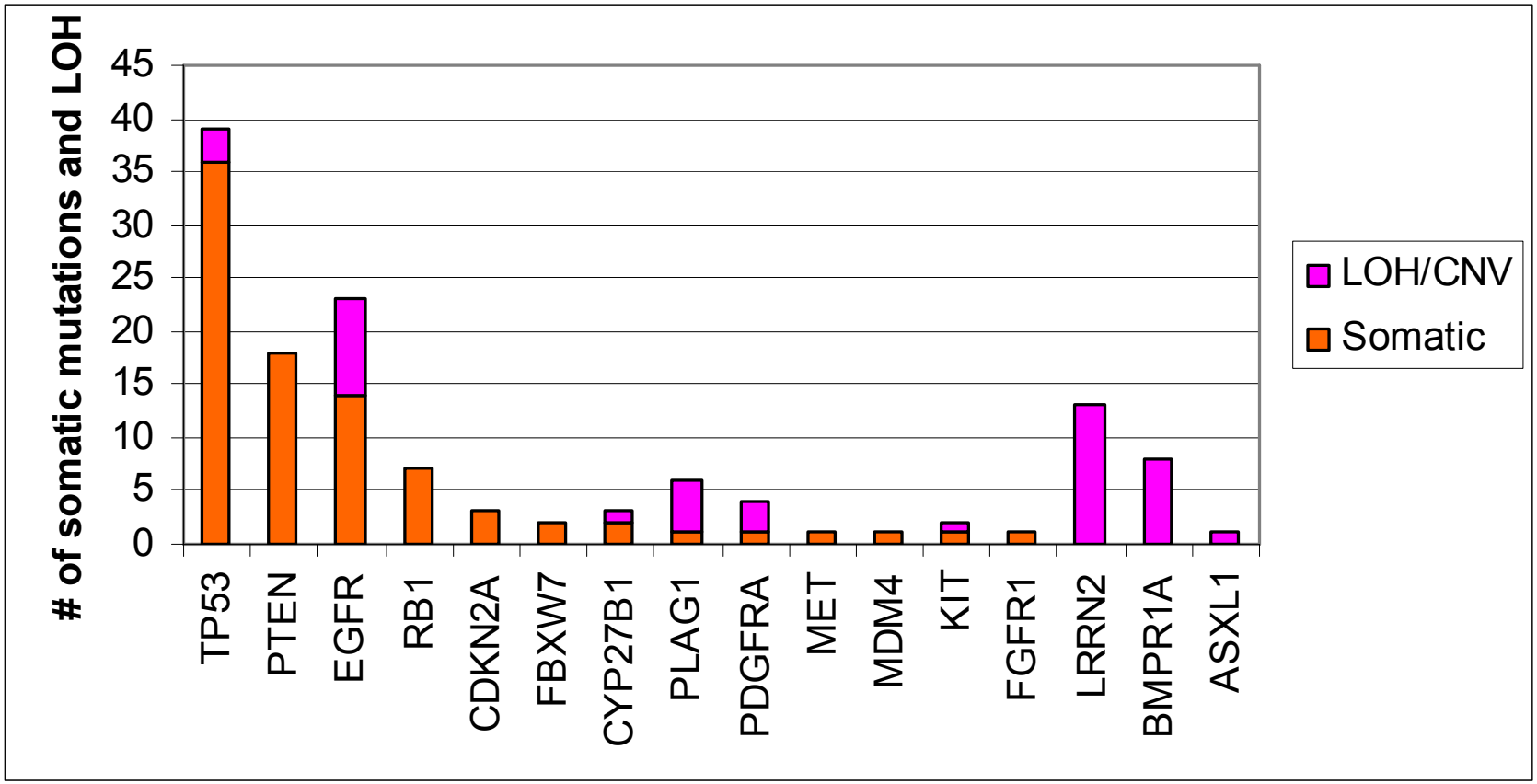
Broad 196 genes: 291 point mutations

Baylor 210 genes: 422 point mutations

Total 605 genes: **1,075 point mutations**

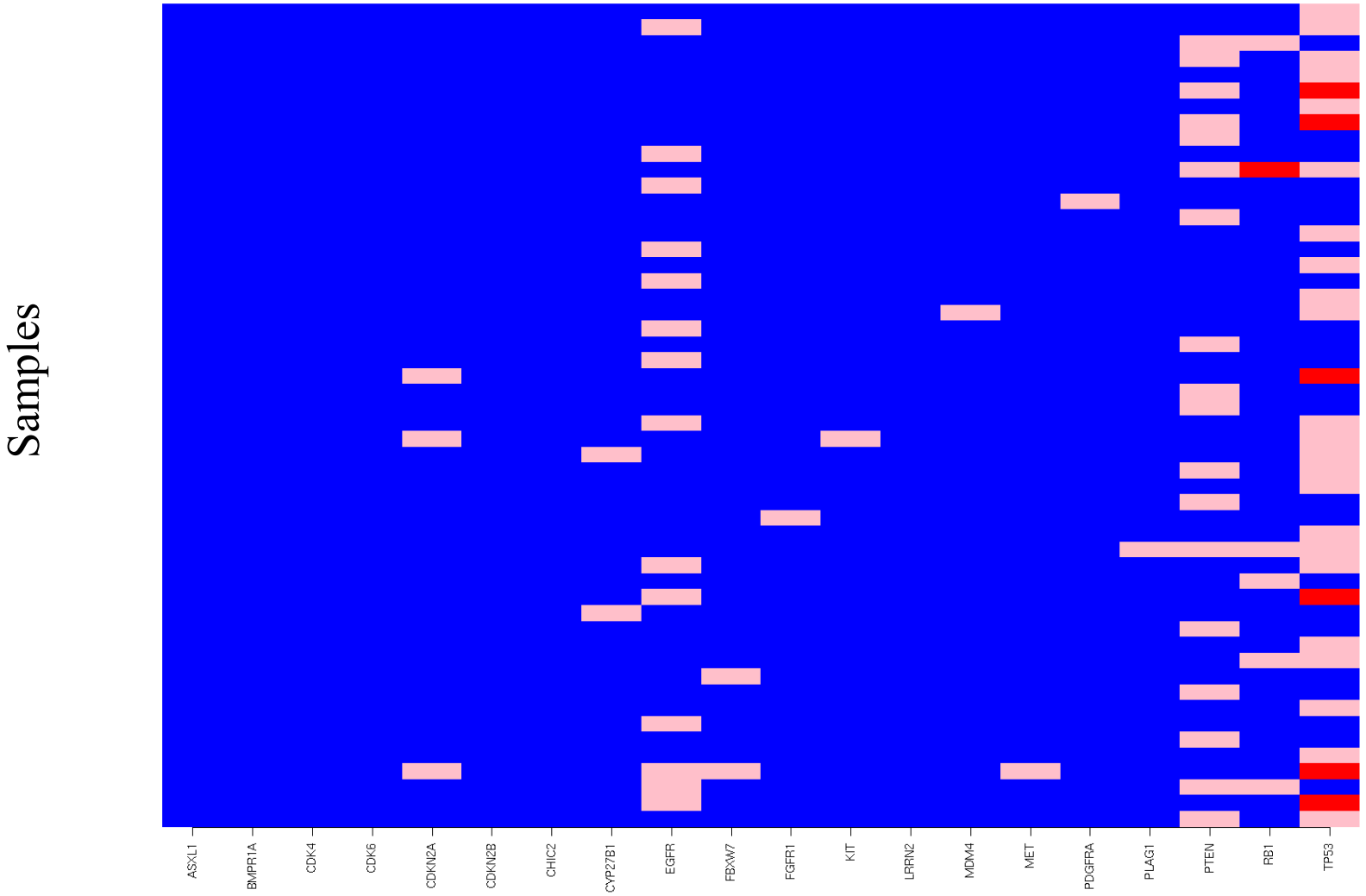
Somatic Mutations and LOH Identified in the Shared 20 Genes

-About 1/3 and 1/5 samples with TP53 and PTEN mutations



- **100 glioblastoma samples and their matched normals have been sequenced**

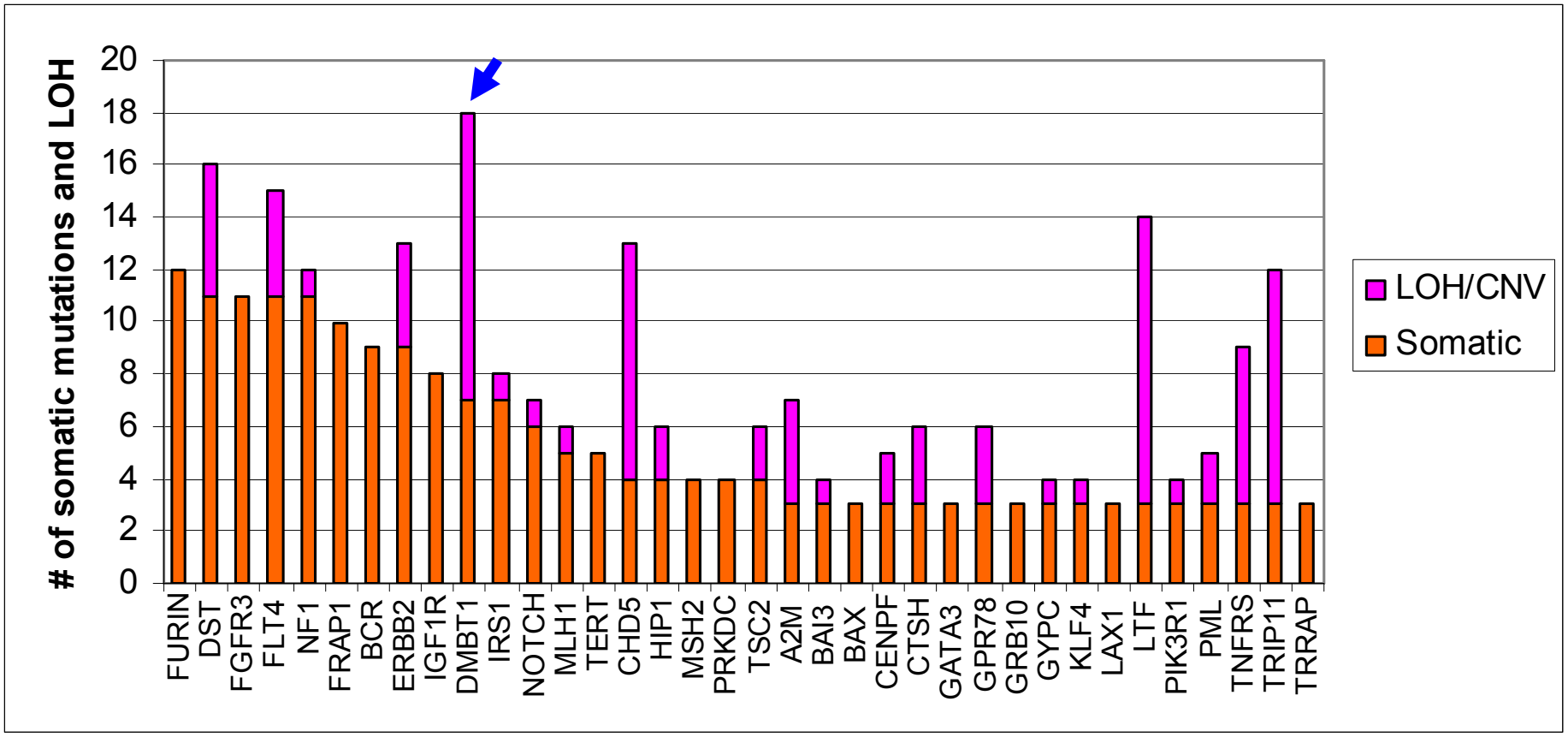
Map of Somatic Mutations in the Shared 20 Genes



Pink: 1 mutation Red: 2 mutations

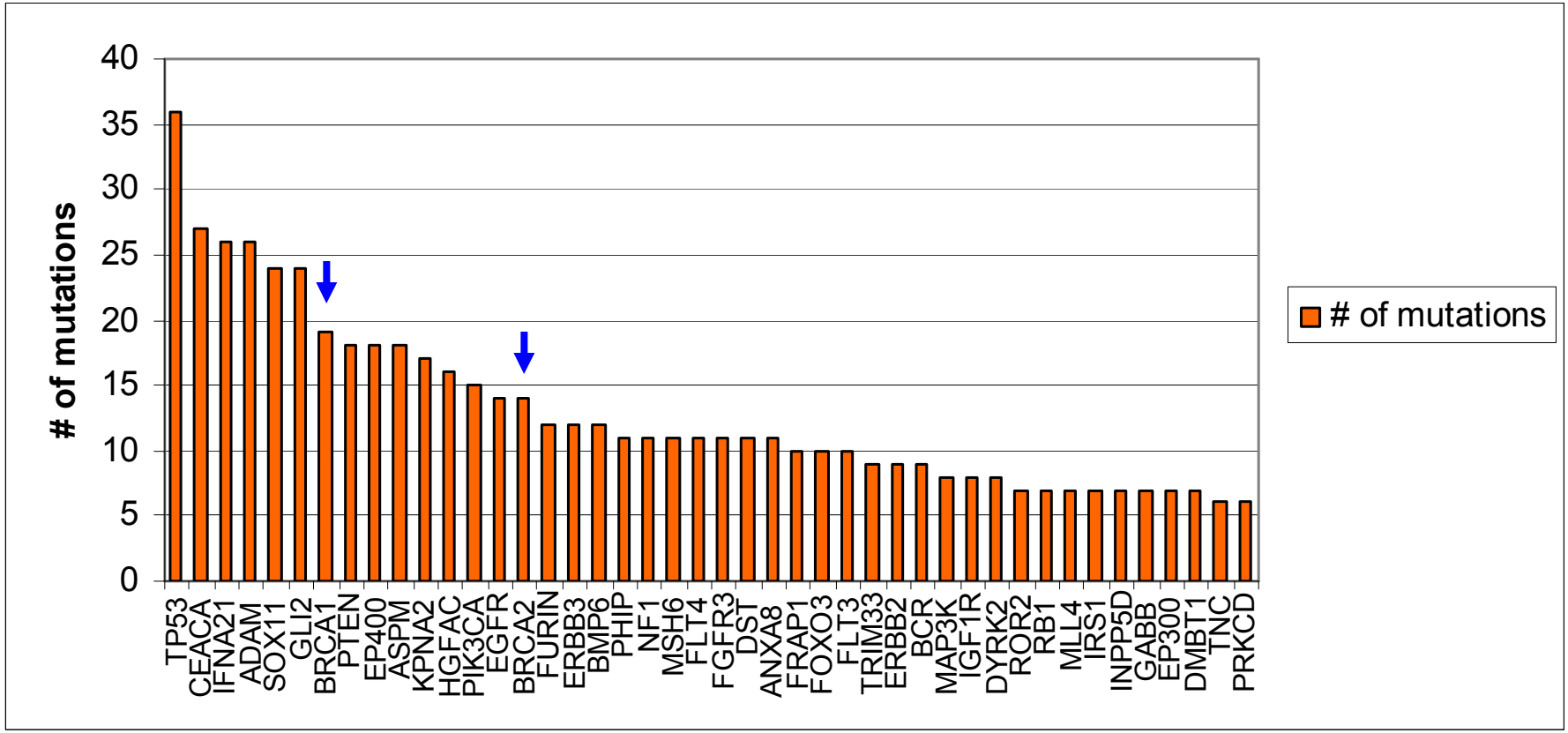
*Indels included

Candidate Somatic Mutations and LOH Identified in 179 WU Center Specific Genes



- **DMBT1(deleted in malignant brain tumors 1): DMBT1 was originally isolated based on its deletion in a medulloblastoma cell line.**
- **100 glioblastoma samples and their matched normals have been sequenced**

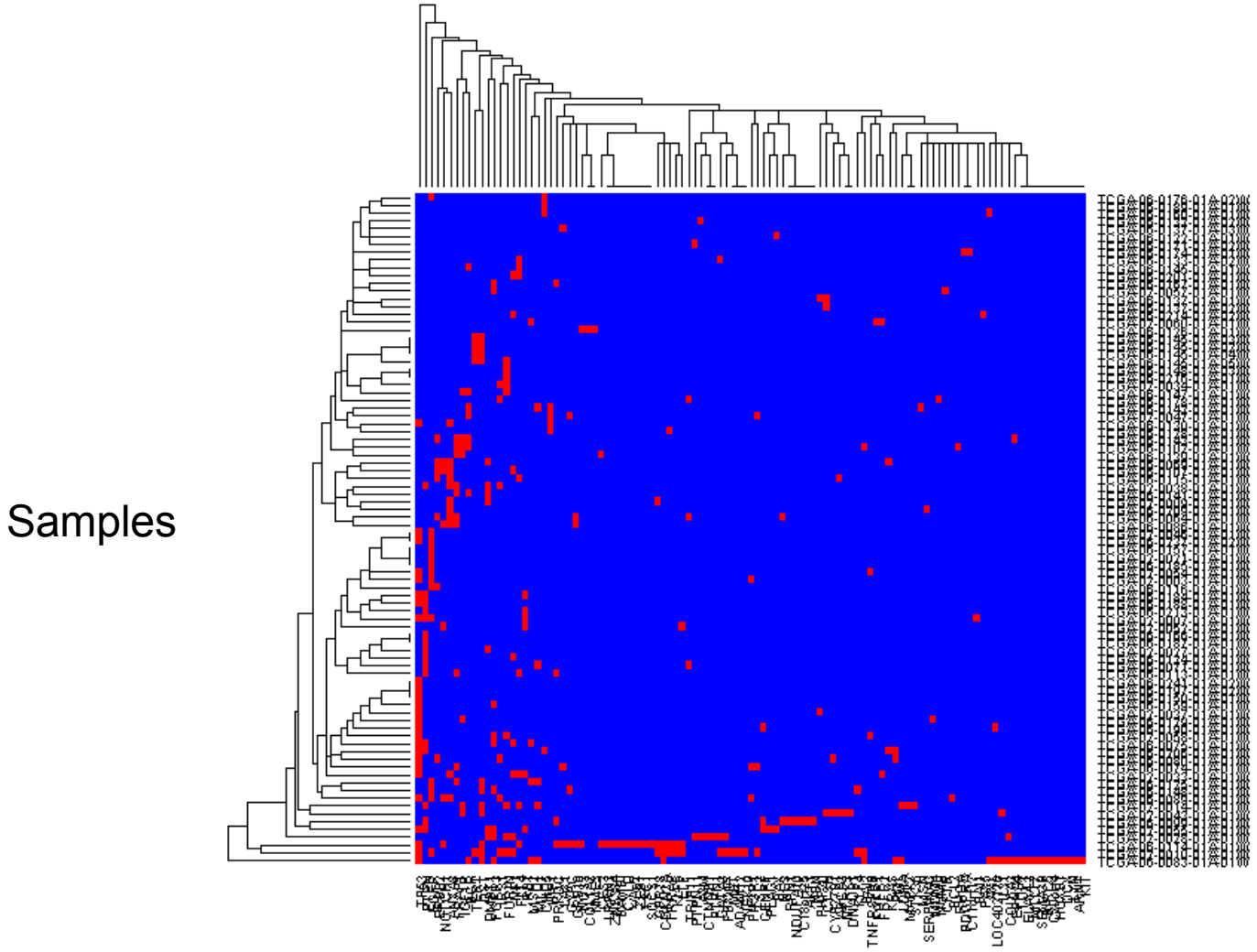
Distribution of **Candidate** Somatic Mutations in 605 Phase I Genes



❖TCGA WU/Broad/Baylor/Shared genes used for this analysis

Two-Way Clustering of TCGA Genes and Samples Based on **Candidate** Mutations

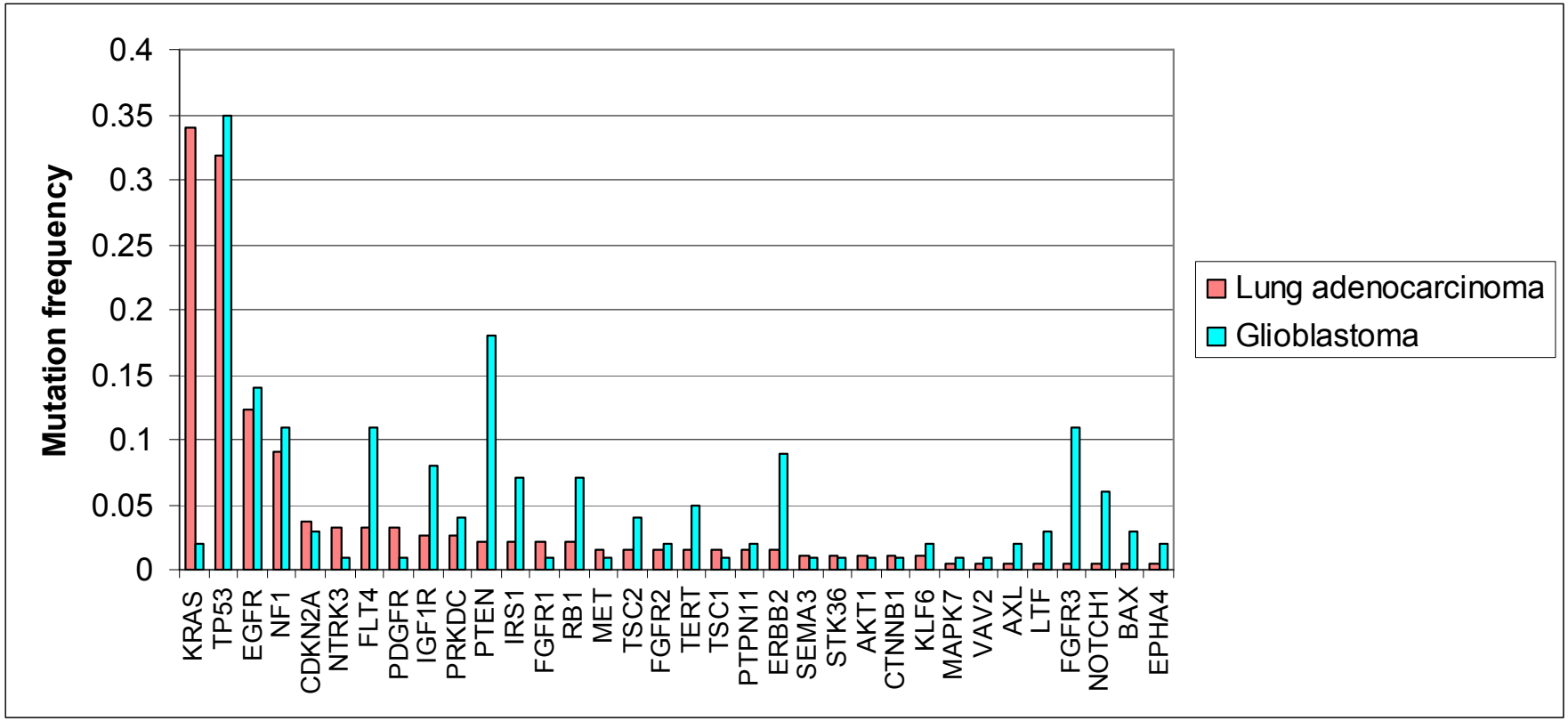
-Do the Sample Clusters Correspond to Any Clinical Subgroups?



❖TCGA WU/Shared genes used for this analysis

TSP & TCGA

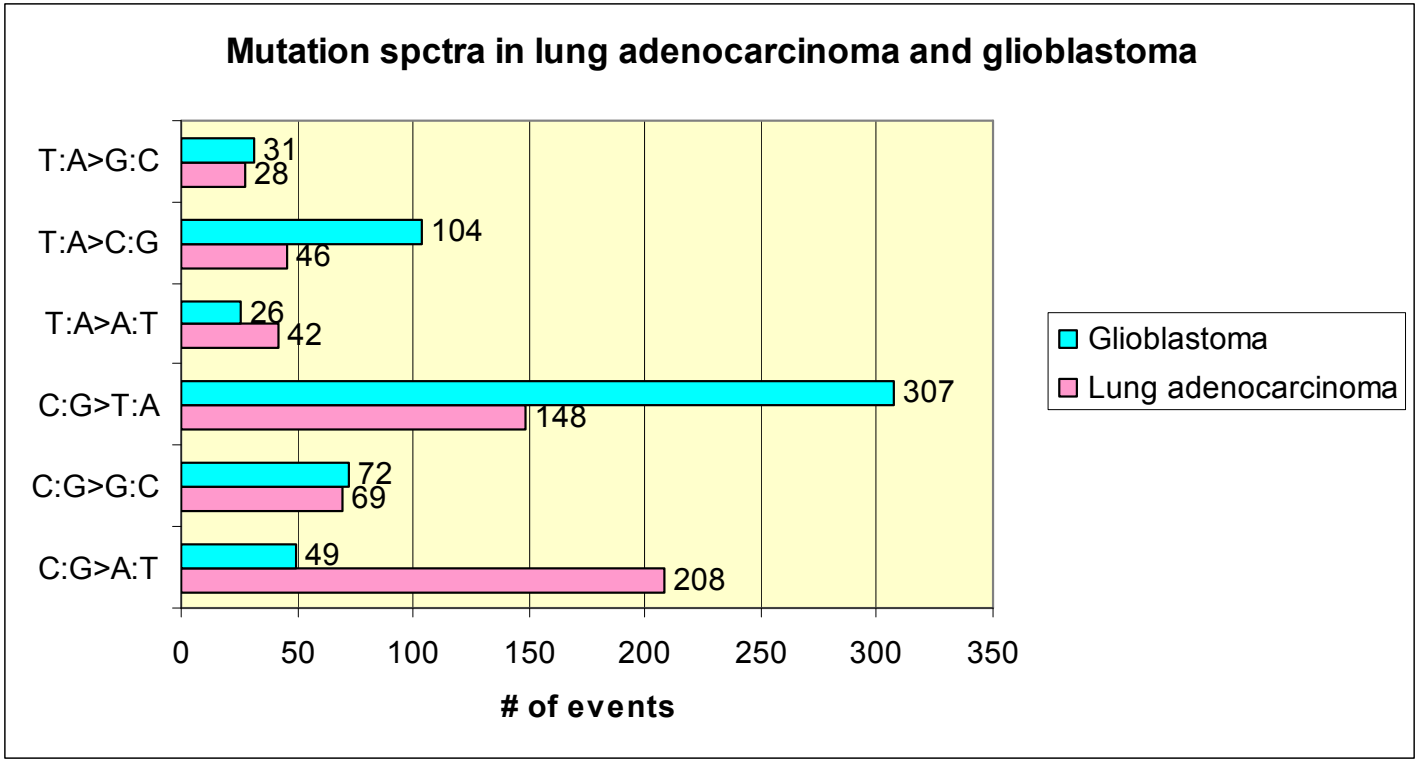
Comparison of Lung Adenocarcinoma and Glioblastoma
-Mutation Rates in Two Cancer Types



- TP53, EGFR, and NF1 have similar mutation rates in both cancer types
- KRAS, FLT4, PTEN, ERBB2, FGFR3, and Notch1 display cancer-type dependent mutation rates

- TSP WU/Broad/Shared genes used for this analysis
- TCGA WU/Shared genes used for this analysis

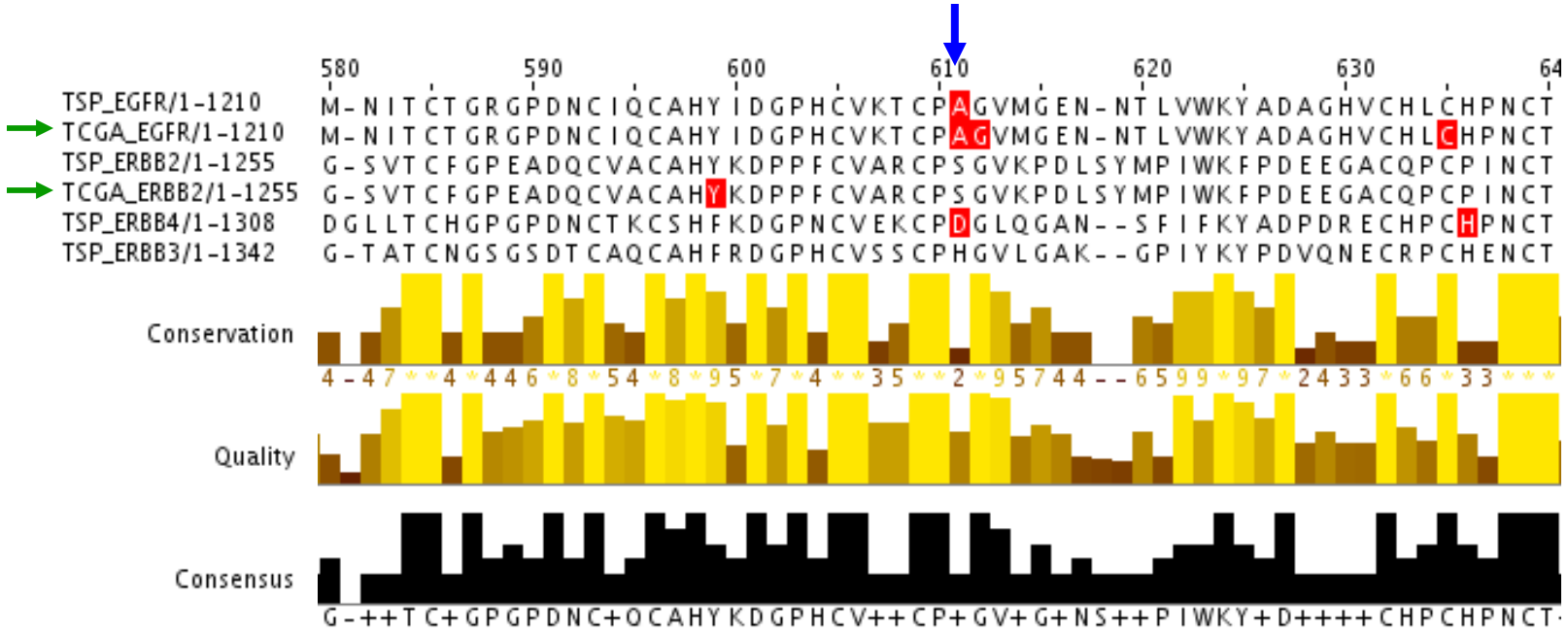
Mutation Signatures in Lung Adenocarcinoma and Glioblastoma



- Glioblastoma: C:G>T:A transition
- Lung adenocarcinoma: C:G>A:T transversion

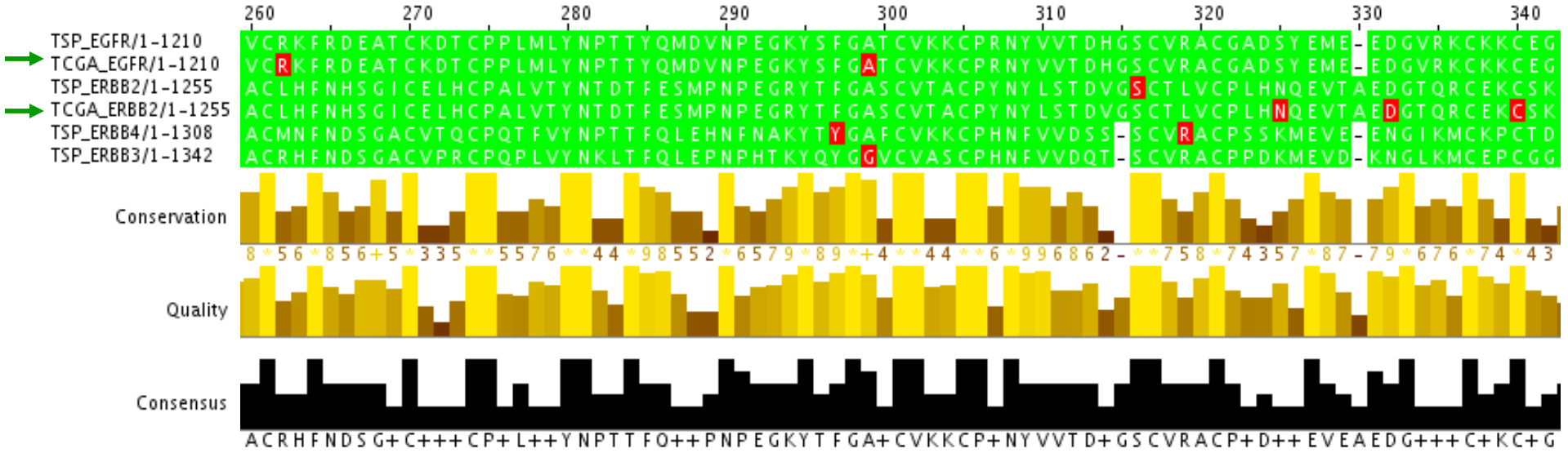
- TSP WU/Broad/Shared genes used for this analysis
- TCGA WU/Shared genes used for this analysis

Common Mutations Found in the Linker Region between Receptor L and Kinase Domains of EGFR Family Members in Lung Adenocarcinoma and Glioblastoma



→ **GBM** Mutations

Mutation Cluster Found in the Furin Like Domain of EGFR, ERBB2, ERBB3, and ERBB4 in Lung Adenocarcinoma and Glioblastoma



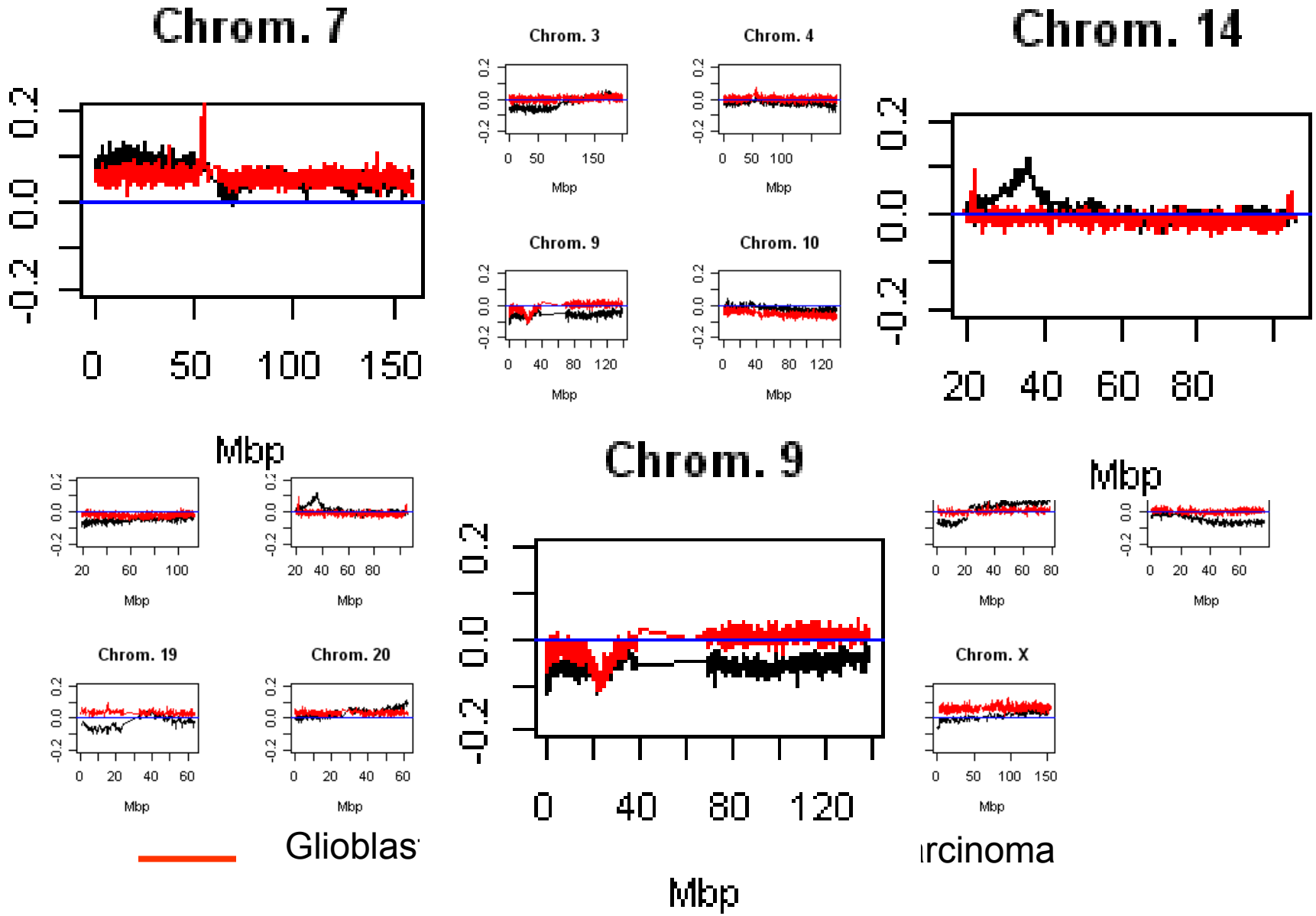
GBM

Furin like domain

Mutations

Landscape of Lung Adenocarcinoma and Glioblastoma

-Copy Number Variation



Summary and Future Directions

- TSP, as a pilot project to explore the feasibility of cancer genome characterization, has facilitated and accelerated TCGA project.

- Highly sensitive and specific mutation discovery and analysis in glioblastoma are underway.

- TSP and TCGA provided the opportunity for comprehensive analysis cross cancer types and cross platforms.

- Next step is unbiased re-sequencing of whole cancer genome using NexGen sequencers.

Acknowledgement

Baylor Human Genome Sequencing Center
David Wheeler
Donna Munzy
George Weinstock
Richard Gibbs

Broad Institute of MIT and Harvard
Carrie Sougnez
Michael Zody
Kristian Cibulskis
Stacey Gabriel
Eric Lander

Dana-Farber
Heidi Greulich
Matthew Meyerson

Washington University

Division of Statistical Genomics
Qunyuan Zhang
Mike Province

Pathology and Immunology
Mark Watson

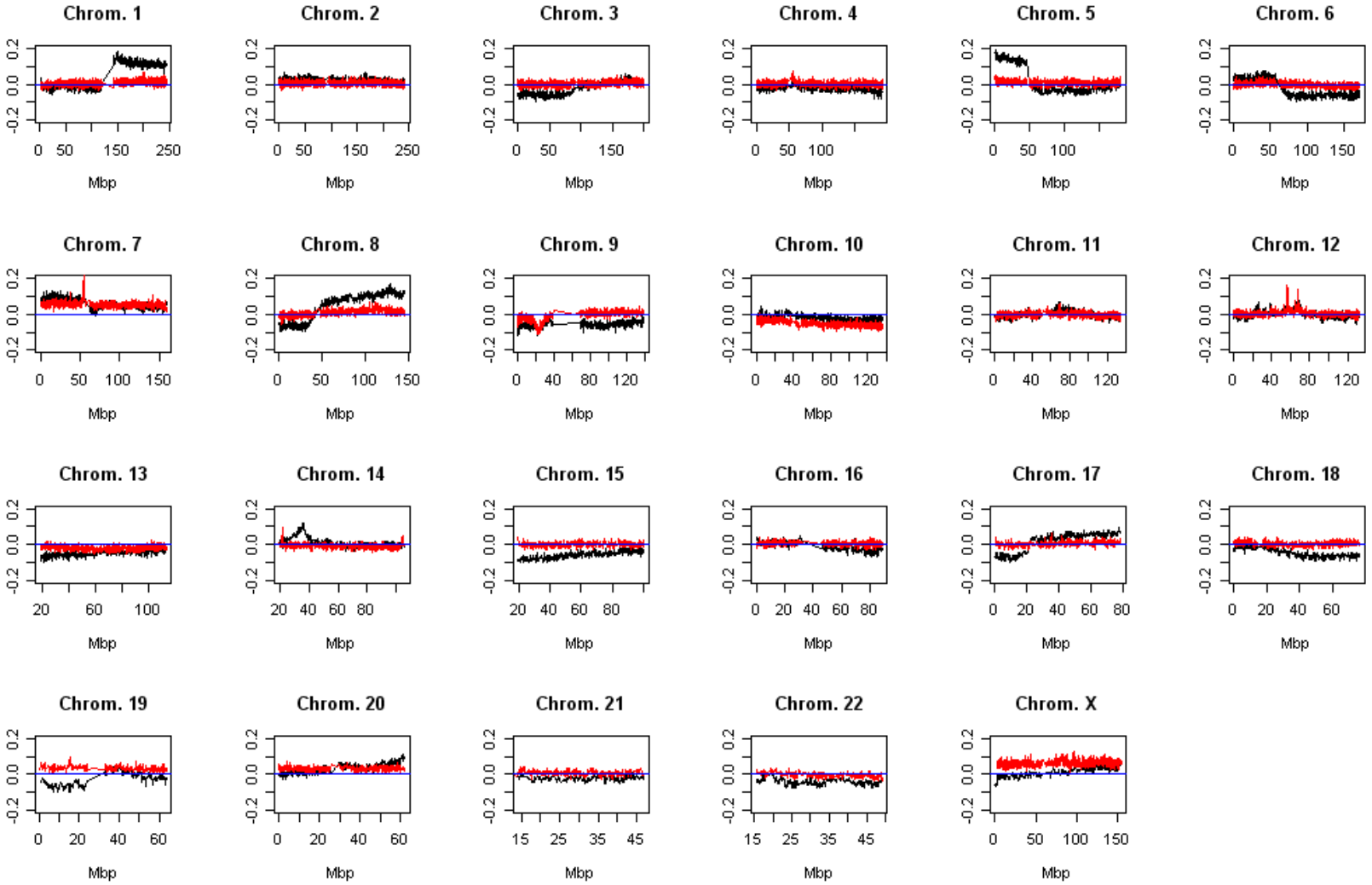
Genome Sequencing center
Mike Mclellan
David Dooling
Tracie Miner
Lucinda Fulton
Elaine Mardis
Rick Wilson

NHGRI & NCI
Bradley Ozenberger

Thanks to everyone involved in TSP and TCGA projects!!!

Landscape of Lung Adenocarcinoma and Glioblastoma

-Copy Number Variation



Glioblastoma

Lung adenocarcinoma