TCGA: Progress and Challenges

Lynda Chin
Belfer Institute for Applied Cancer Science
Dana-Farber Cancer Institute
Harvard Medical School
Broad Institute

September 8, 2010

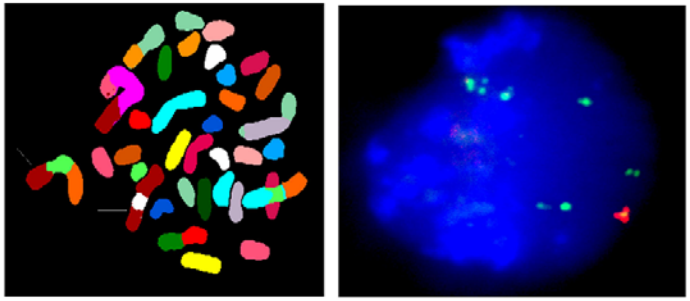
Goals of cancer medicine and the promise of Cancer Genomics

- Prevention
  - Understanding the underlying etiology → strategy
- Detection
  - Identify risk alleles / genomic events for screening
- Intervention
  - Stratify high vs low risk patients to treat or not
  - Identify new therapeutic targets for drug discovery
  - Inform selection of the right patient for the right drug
  - Define combination / co-extinction strategies
  - Understand resistance mechanisms

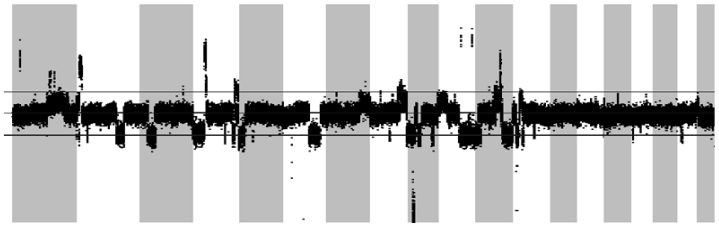
Multi-dimensional Cancer Genomics

THE CANCER GENOME ATLAS

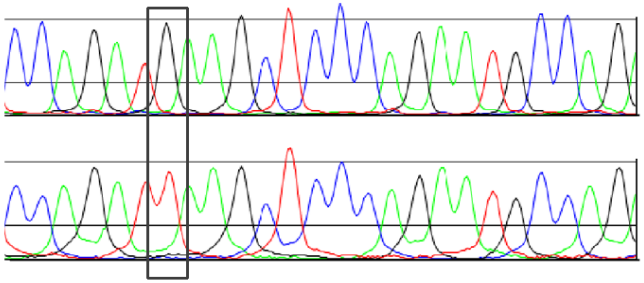
Aneuploidy; Re-arrangement; Translocation



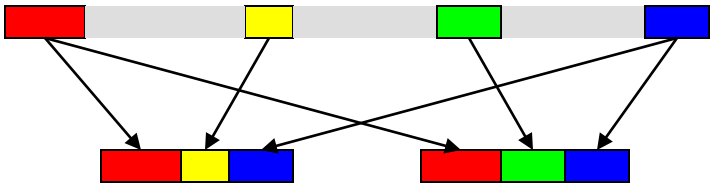
Copy number aberrations



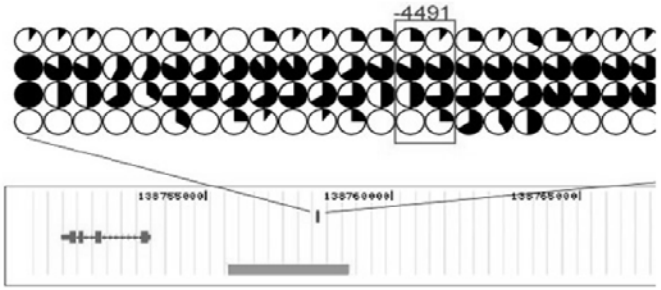
Somatic mutations



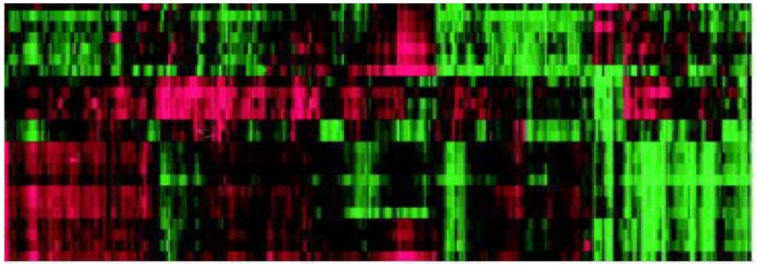
Gene Splicing Alterations



Methylation or histone modification



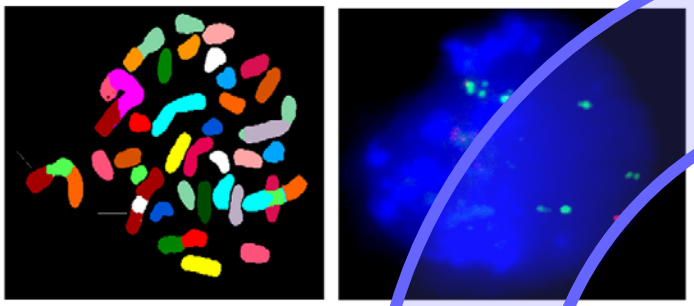
Altered expression



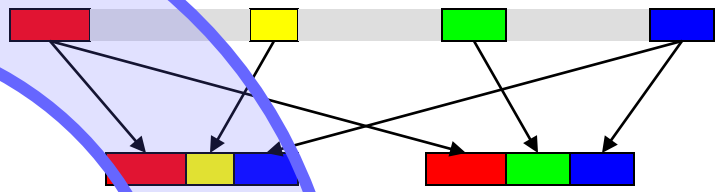
TCGA Pilot (2006 – 2009)

THE CANCER GENOME ATLAS

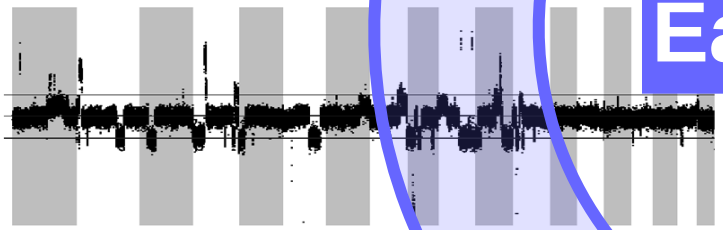
Aneuploidy; Re-arrangement; Translocation



Gene Splicing Alterations

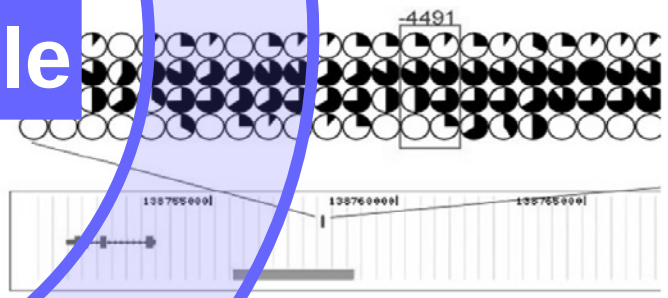


Copy number aberrations

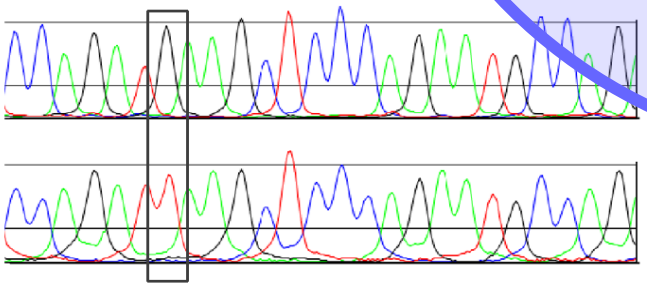


Each Sample

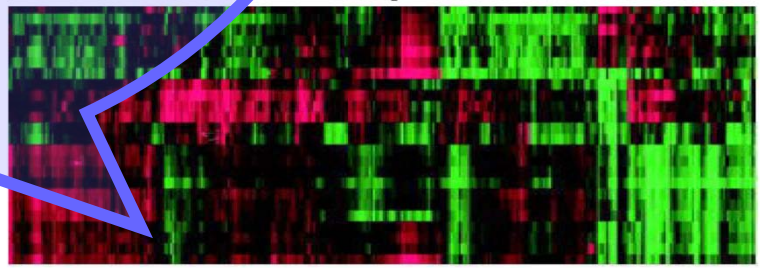
Methylation or histone modification



Somatic mutations



Altered expression



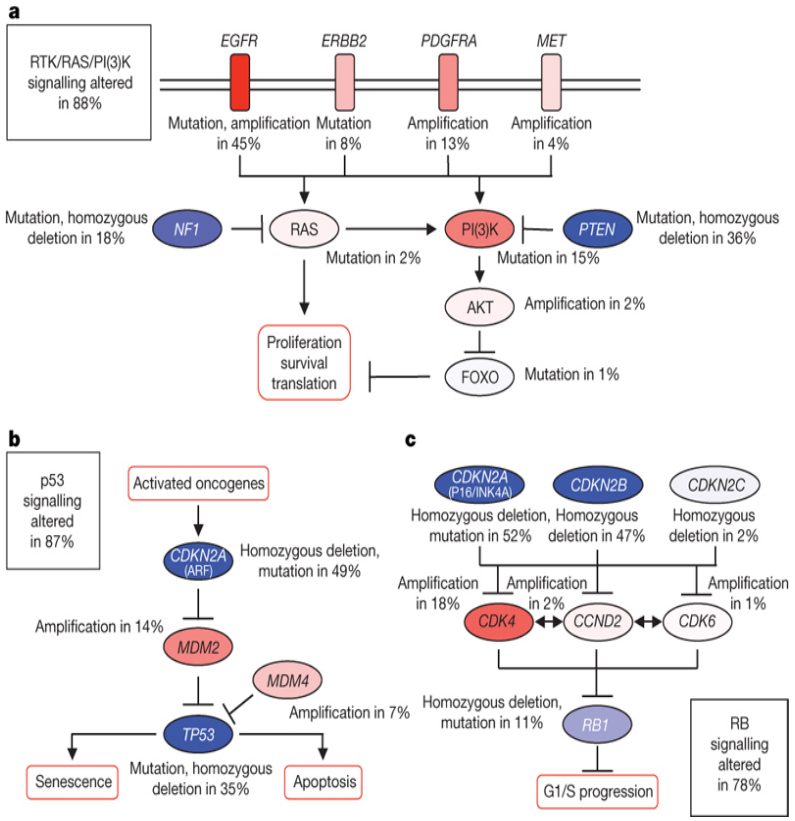
doi:10.1038/nature07385

nature

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

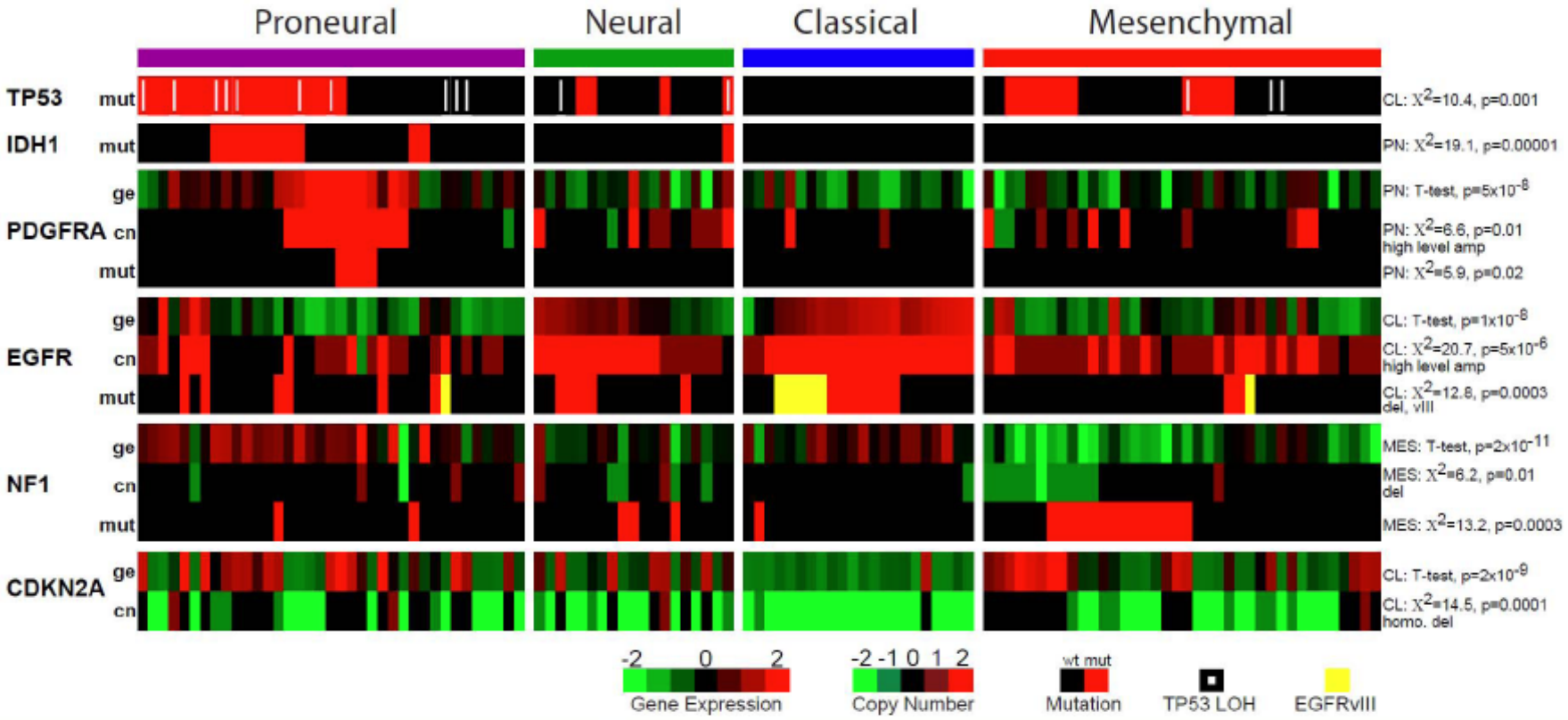
- A Reference GBM cancer genome
  - PIK3R1 mutation is frequent in GBM
  - NF1 is involved in sporadic GBM in human
  - TP53 is commonly mutated in primary GBM
- Unanticipated discoveries..
  - Hypothesis on a possible resistance mechanism to temozolomide (TMZ)
- Integrative analyses → Pathway knowledge



Cancer Cell
Article

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

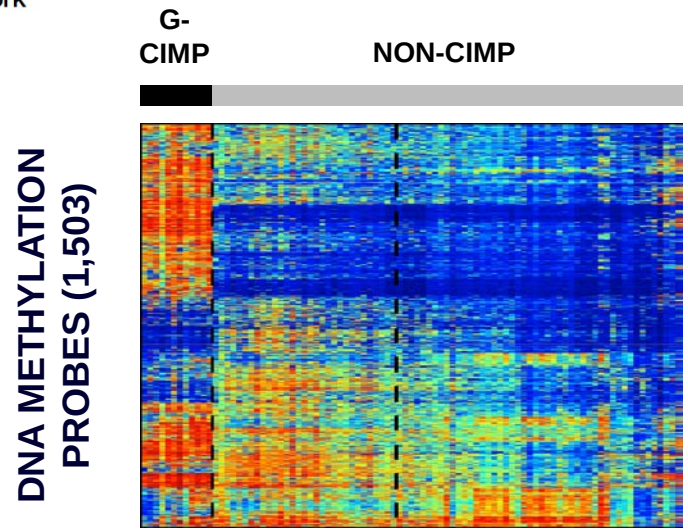
Roel G.W. Verhaak,[1,2,17] Katherine A. Hoadley,[3,4,17] Elizabeth Purdom,[7] Victoria Wang,[8] Yuan Qi,[4,5] Matthew D. Wilkerson,[4,5] C. Ryan Miller,[4,6] Li Ding,[9] Todd Golub,[1,10] Jill P. Mesirov,[1] Gabriele Alexe,[1] Michael Lawrence,[1,2] Michael O'Kelly,[1,2] Pablo Tamayo,[1] Barbara A. Weir,[1,2] Stacey Gabriel,[1] Wendy Winckler,[1,2] Supriya Gupta,[1] Lakshmi Jakkula,[11] Heidi S. Feiler,[11] J. Graeme Hodgson,[12] C. David James,[12] Jann N. Sarkaria,[13] Cameron Brennan,[14] Ari Kahn,[15] Paul T. Spellman,[11] Richard K. Wilson,[9] Terence P. Speed,[7,16] Joe W. Gray,[11] Matthew Meyerson,[1,2] Gad Getz,[1] Charles M. Perou,[3,4,8] D. Neil Hayes,[4,5,*] and The Cancer Genome Atlas Research Network

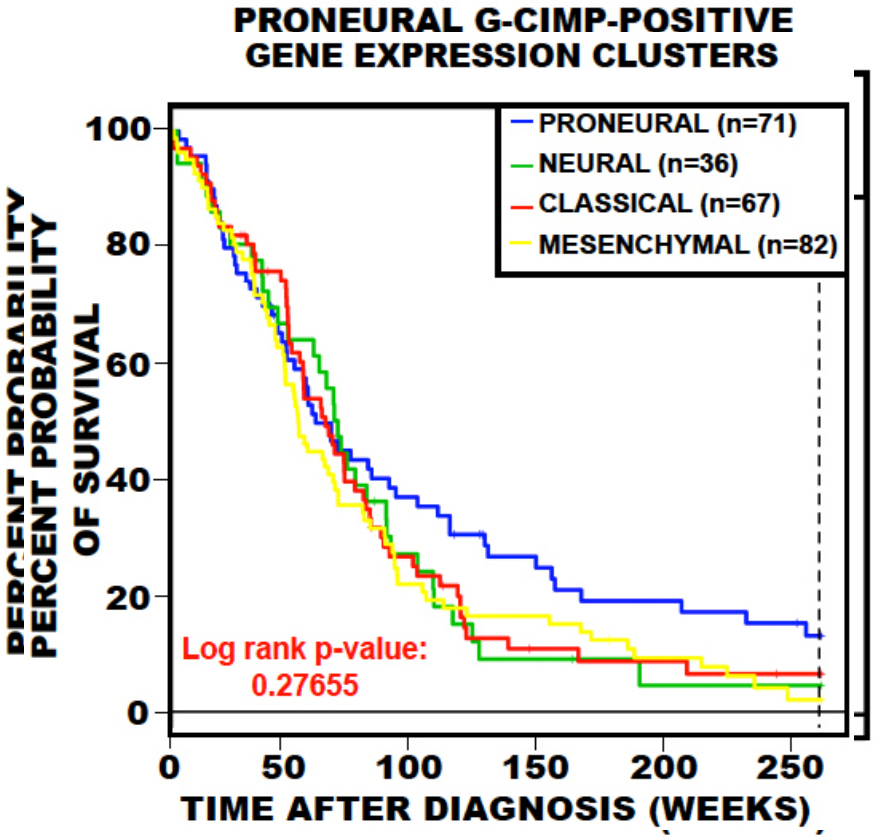


Cancer Cell
Article

Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma

Houtan Noushmehr,[1,13] Daniel J. Weisenberger,[1,13] Kristin Diefes,[2,13] Heidi S. Phillips,[3] Kanan Pujara,[3] Benjamin P. Berman,[1] Fei Pan,[1] Christopher E. Pelloski,[4] Erik P. Sulman,[4] Krishna P. Bhat,[2] Roel G.W. Verhaak,[5,6] Katherine A. Hoadley,[7,8] D. Neil Hayes,[7,8] Charles M. Perou,[7,8] Heather K. Schmidt,[9] Li Ding,[9] Richard K. Wilson,[9] David Van Den Berg,[1] Hui Shen,[1] Henrik Bengtsson,[10] Pierre Neuvial,[10] Leslie M. Cope,[11] Jonathan Buckley,[1,12] James G. Herman,[11] Stephen B. Baylin,[11] Peter W. Laird,[1,14,*] Kenneth Aldape,[2,14] and The Cancer Genome Atlas Research Network

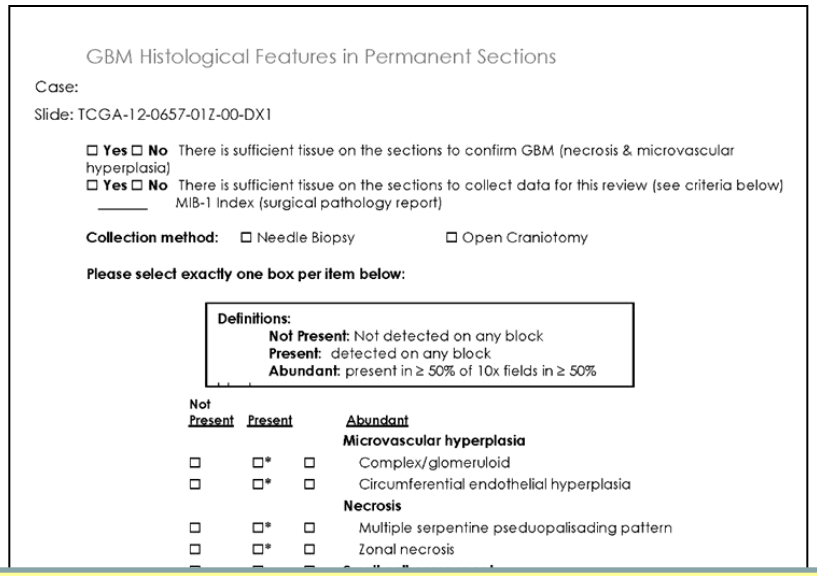




- Occurs in Younger Patients
- Is a Subset of Proneural Expression Subtype
- Is Associated with Better Survival
- Is More Frequent in Low-Grade Gliomas
- Is Not Associated with *MGMT* Methylation
- Is Tightly Linked to *IDH1* Mutation

Noushmehr et al. (2010) Cancer Cell, Online

Clinicopathological correlation…



GBM Histological Features in Permanent Sections

Case:

Slide: TCGA-12-0657-01Z-00-DX1

☐ **Yes** ☐ **No** There is sufficient tissue on the sections to confirm GBM (necrosis & microvascular hyperplasia)

☐ **Yes** ☐ **No** There is sufficient tissue on the sections to collect data for this review (see criteria below)

_____ MIB-1 Index (surgical pathology report)

Collection method: ☐ Needle Biopsy ☐ Open Craniotomy

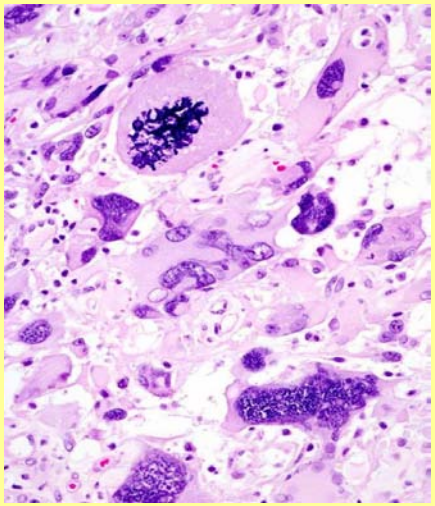
Please select exactly one box per item below:

Definitions:
Not Present: Not detected on any block
Present: detected on any block
Abundant: present in ≥ 50% of 10x fields in ≥ 50%

Not Present	Present	Abundant	
			Microvascular hyperplasia
☐	☐*	☐	Complex/glomeruloid
☐	☐*	☐	Circumferential endothelial hyperplasia
			Necrosis
☐	☐*	☐	Multiple serpentine pseudopalisading pattern
☐	☐*	☐	Zonal necrosis

Consensus path review on digital H&E images

Daniel Brat
Scott Vandenberg
Roger McLendon
David Louis
Norm Lehman
Mark Cohen
Ryan Miller
Matt Schniederjan



Giant Cells	p53-wt	p53 mut/del
0	80%	20%
1+	33%	67%
2+	0%	100%

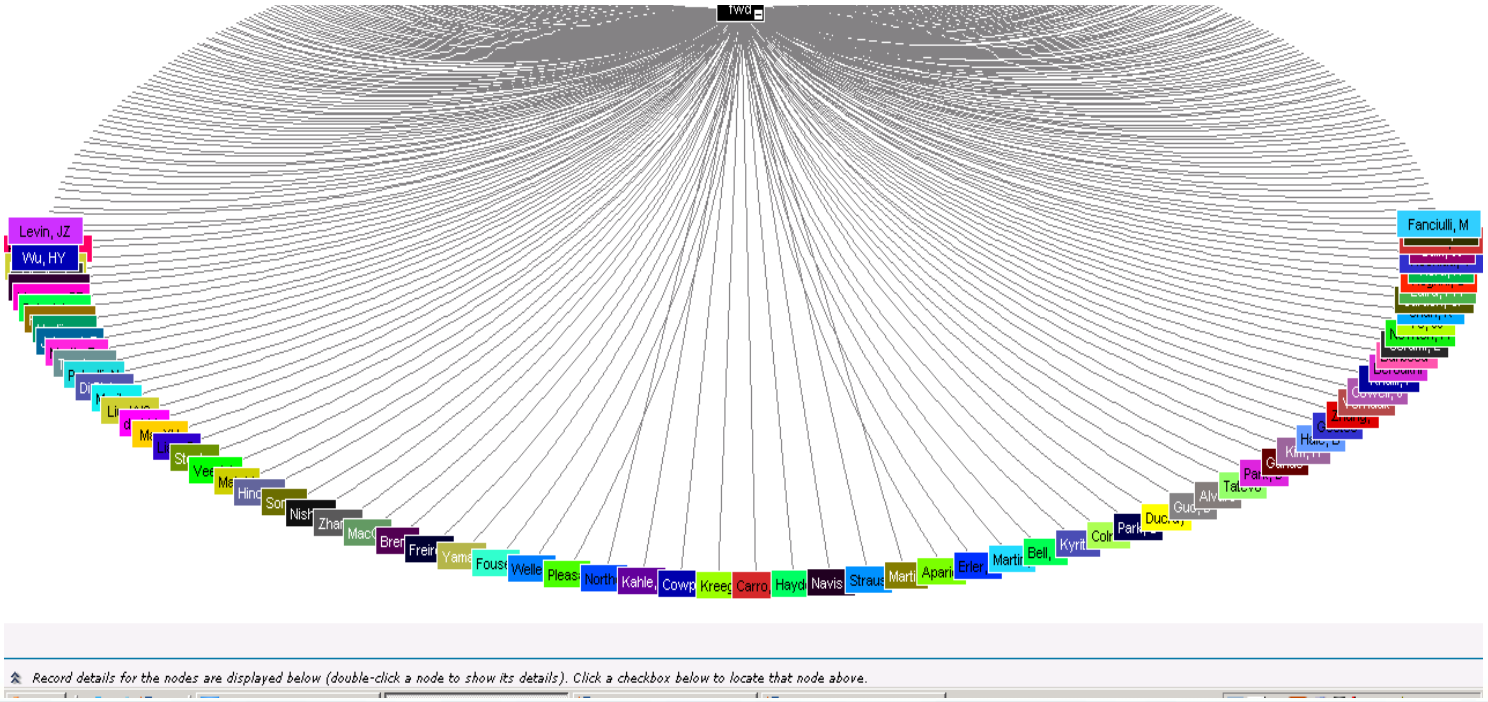
p<= 6.7 e-05

Giant Cells	P53 pathway intact	p53 pathway altered
0	75%	25%
1+	14%	85.7%
2+	0%	100%

p<= 1.4 e-06

Enabling resource

- Citation in 225 publications
  - Comparison with mouse models
  - Novel gene discovery and pathway analyses
  - Analysis of germline genetics
  - Novel computational algorithm development
  - In silico correlation studies



Conclusions from the pilot

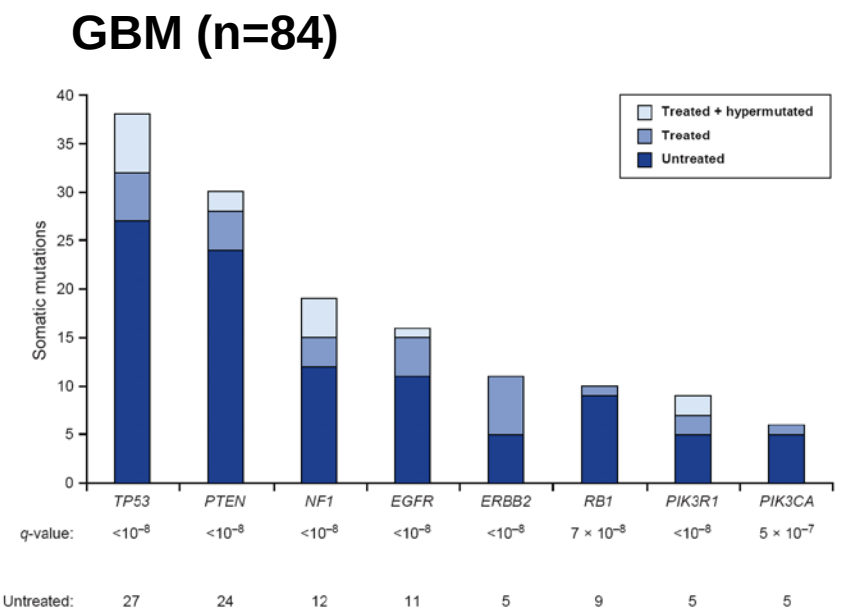
- Cancer genome is highly complex and heterogeneous
  - Technologies can detect the signals above the noises
- There are new discoveries to be made
  - Detect known and discover unknown genes
  - Discovery of novel subclass, e.g. G-CIMP
- Multi-dimensional analyses enable integrative analyses
  - Pathway → Network view → translational potential
- Unbiased approach generates unanticipated hypotheses
  - Mechanism for TMZ resistance
- Reference-quality data with stringent QC as an enabling resource
  - GBM dataset has been used/referenced in > 225 publications
- The acquisition of large cohorts of high-quality clinically annotated tumor samples is critical but extremely challenging
  - Investment in biospecimen banking / infrastructure

Unique Challenges of TCGA

- **Reference = Complete + Quality**
  - Quality: samples → biomolecules → data → analyses
  - Complete: Multi-dimensionality; global assays
  - Complete: sufficiently powered sample size

What is the power of a discovery set of 21 samples? (Wood et al.)

- We took 100,000 subsets of 21 samples out of the 84 (non-hyper mutated GBM samples from our paper) and calculated the frequency that each of the 8 significant genes would have been detected as significant



Gene	Detected in fraction
TP53	97.5%
PTEN	95.6%
EGFR	70.9%
NF1	69.5%
RB1	48.9%
ERBB2	51.4%
PIK3R1	33.3%
PIK3CA	21.8%

- Stage 1 = 200 Discovery set
  - 20 whole genomes + 180 whole exome
- Stage 2 = 300 Extension validation set
  - targeted sequencing of ~3000-6000 most significant genes

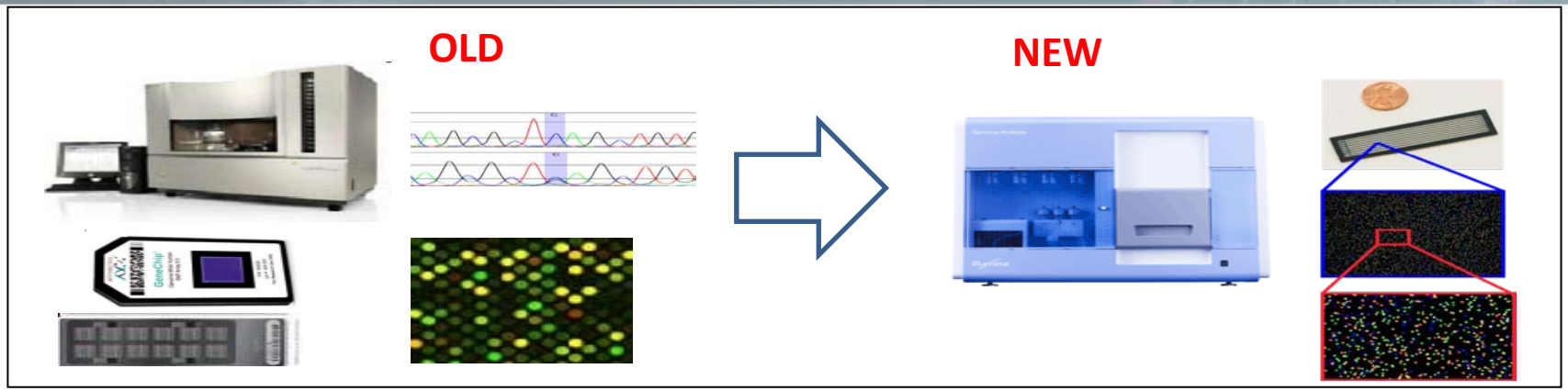
>80% power to detect 3% frequency event

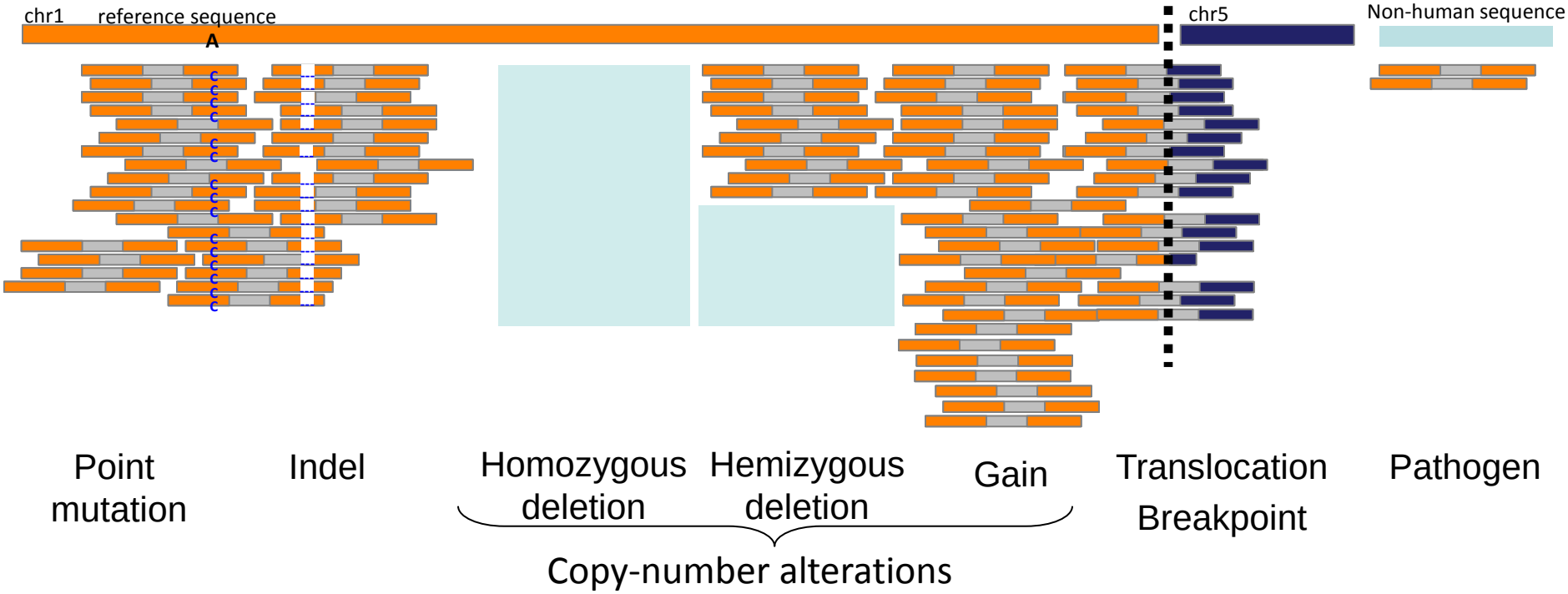
- **Reference = Complete + Quality**
  - Quality: samples → biomolecules → data → analyses
  - Complete: Multi-dimensionality; global assays
  - Complete: sufficiently powered sample size

→Transformative Technology Revolution

Massively Parallel Sequencing

Massively Parallel Sequencing





Example of a cancer genome

THE CANCER GENOME ATLAS

GLIOBLASTOMA

Coverage(T/N)	Callable	Purity	Ploidy
30x / 30x	85%	65%	5.5

Name	TCGA-06-0188
Alias	GBM-0188
Issued By	Broad Institute
Issue Date	July 8, 2009

Point Mutations

Rate/Mb	1.21
Total	3164
Coding	27

MIS	23
STOP	---
INDEL	1

HIGHLIGHTS

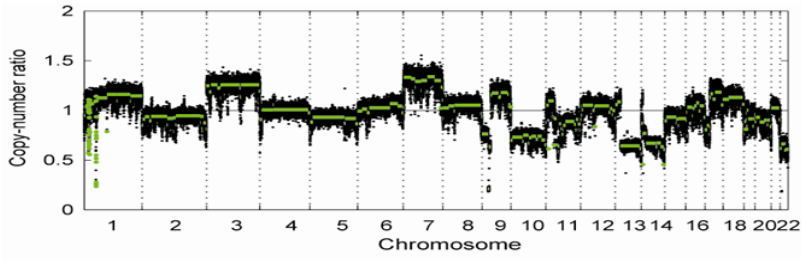
TP53	DNP Splice_site	Tumor suppressor
PTPRB	Missense	Tumor suppressor family member
PTEN	Indel	Tumor suppressor
TNC	Missense	Glioma associated extracellular matrix antigen. Involved in migration of neurons and axons during development

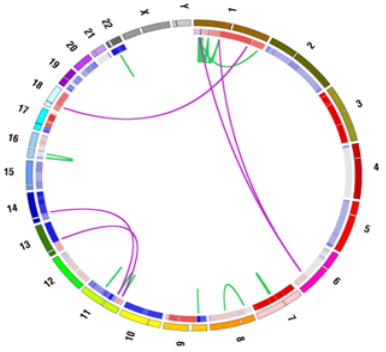
Chr. Aberrations

CNA Breaks	---
TX-Inter	6
TX-Intra	84

HIGHLIGHTS

Major rearrangements in chr1 including CDKN2C and FAF1



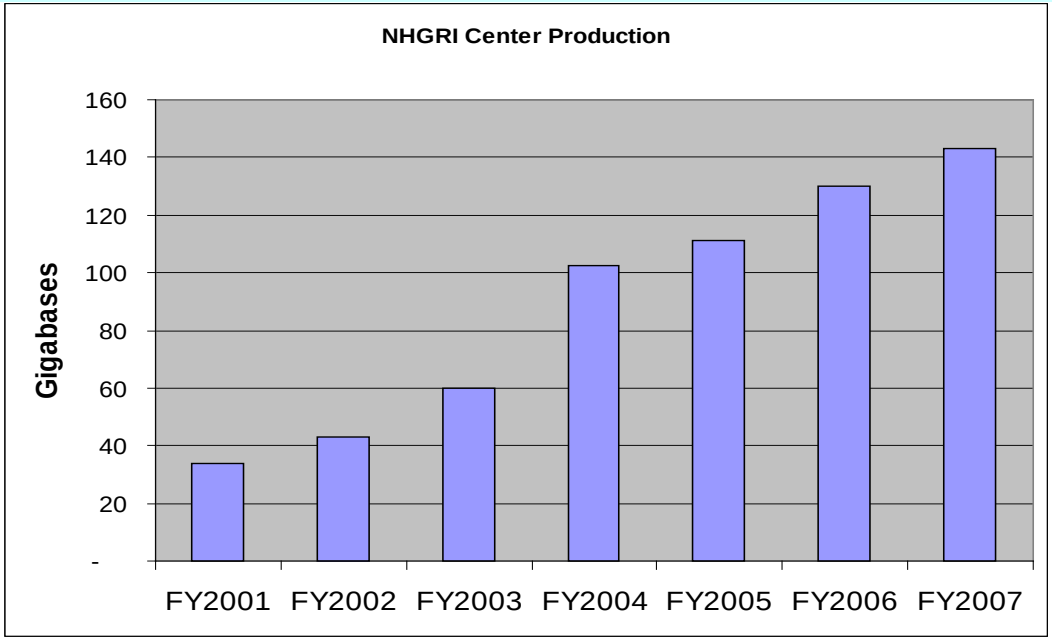


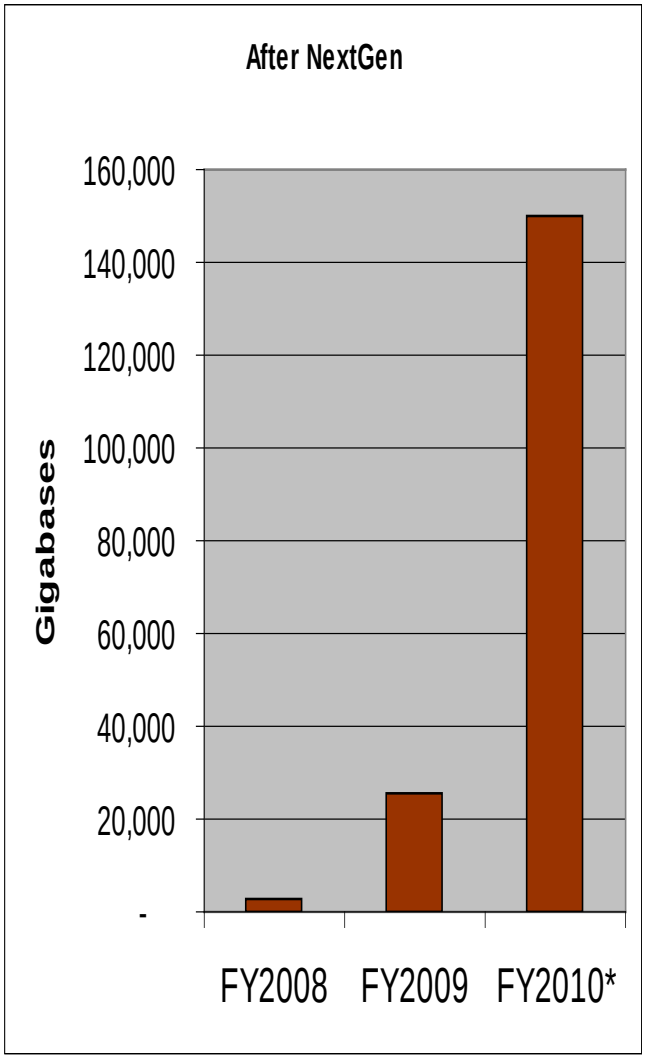
Scale of Growth is unprecedented

THE CANCER GENOME ATLAS

Examples of technical challenges:

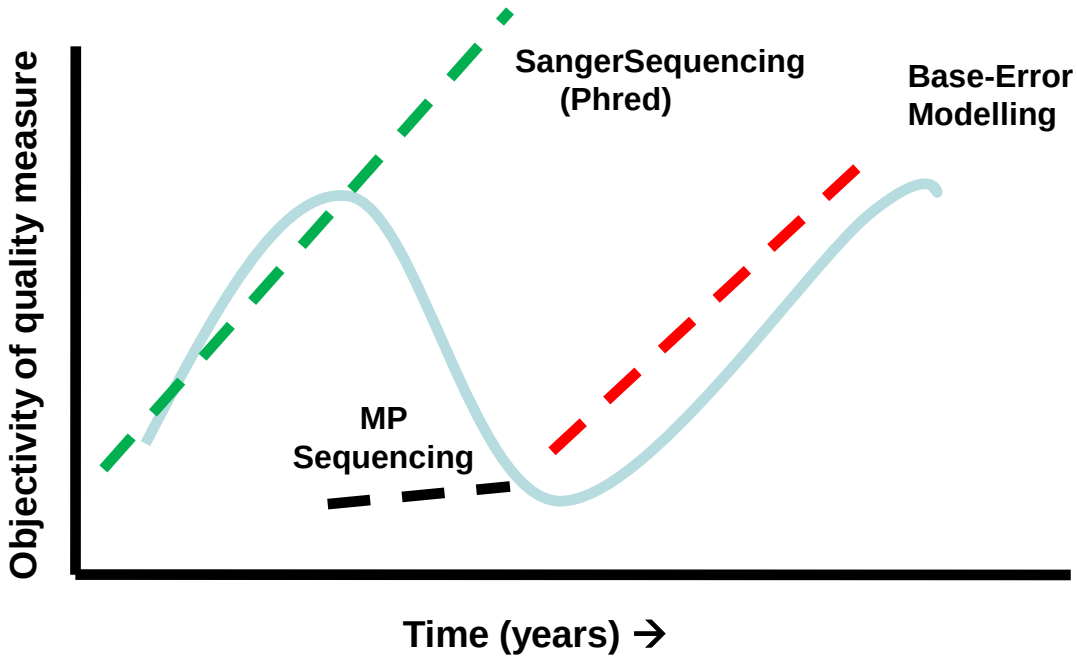
- IT infrastructure
- Optimization of library generation
- Input requirement
- Alignment to genome
- Variance calling algorithms





Validation Challenges

- Currently every variant must be 'validated'
  - For a whole genome, this is thousands of variants and the cost can dwarf discovery cost,
  - Focus on coding regions – still hundreds per tumor type
  - Need to improve error models and practicality of mass-validation



- OVCa MS: all 20,398 somatic variants are being (already) validated in a 2nd assay (by genotyping or repeat sequencing) in all samples

TCGA IS GENERATING NEW KNOWLEDGE

In the midst of a technology revolution

Pattern of somatic genomic alterations

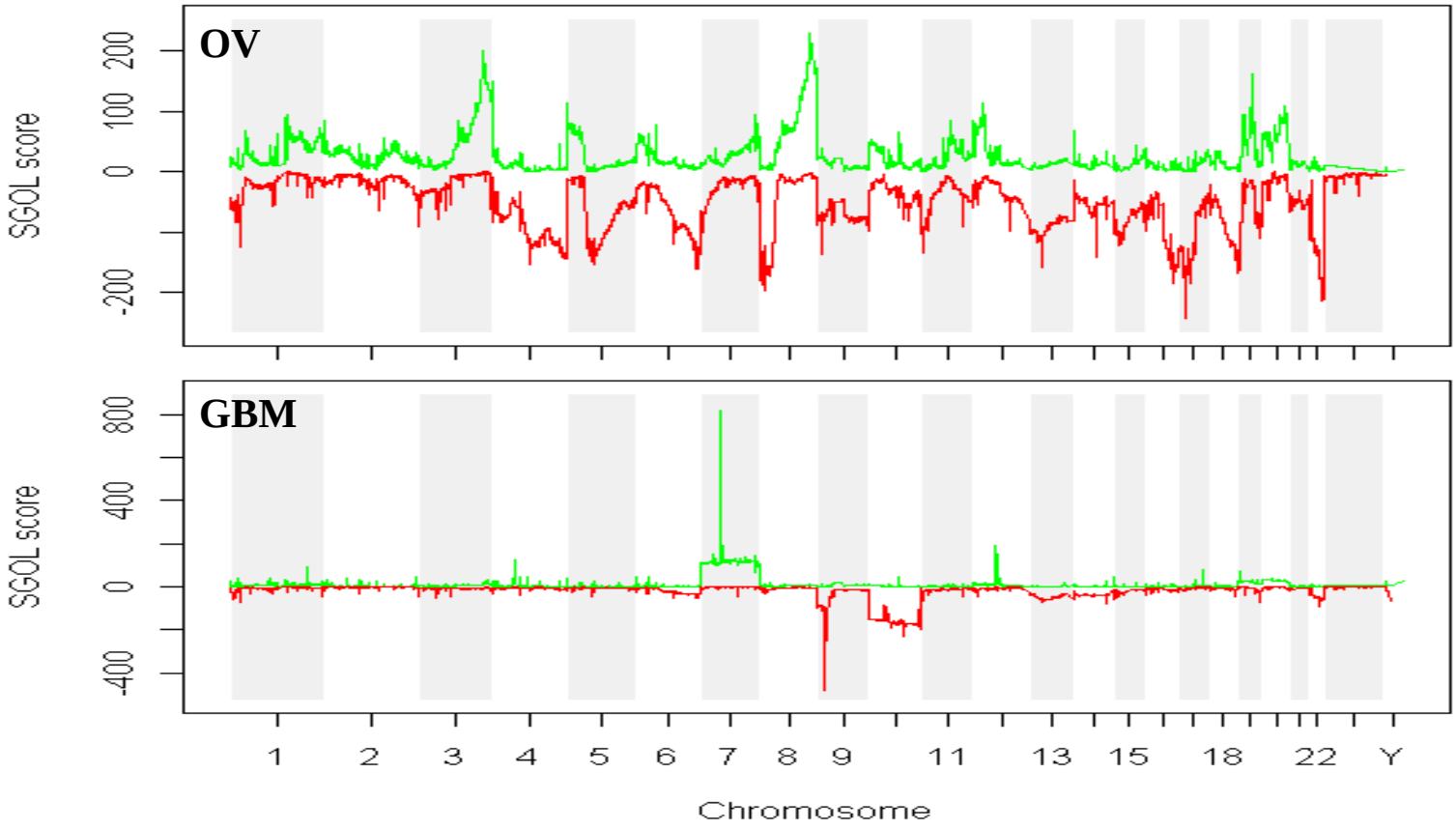
THE CANCER GENOME ATLAS

Significantly mutated genes in serous ovarian cancer (n=316)	
Gene	**# of Mutations**
TP53	277
FAT3	19
CSMD3	18
NF1	14
BRCA1	10
RB1	9
CDK12	9
BRCA2	9
RB1CC1	7
GABRA6	6
TACC3	5

- *TP53* was mutated in 96.5%
- *BRCA1/2* were mutated in 21% of tumors due to germline (9%/6%) or somatic (3%) mutations.
- Other significantly mutated genes in serous OvCa were present in only 1-6% of tumors.

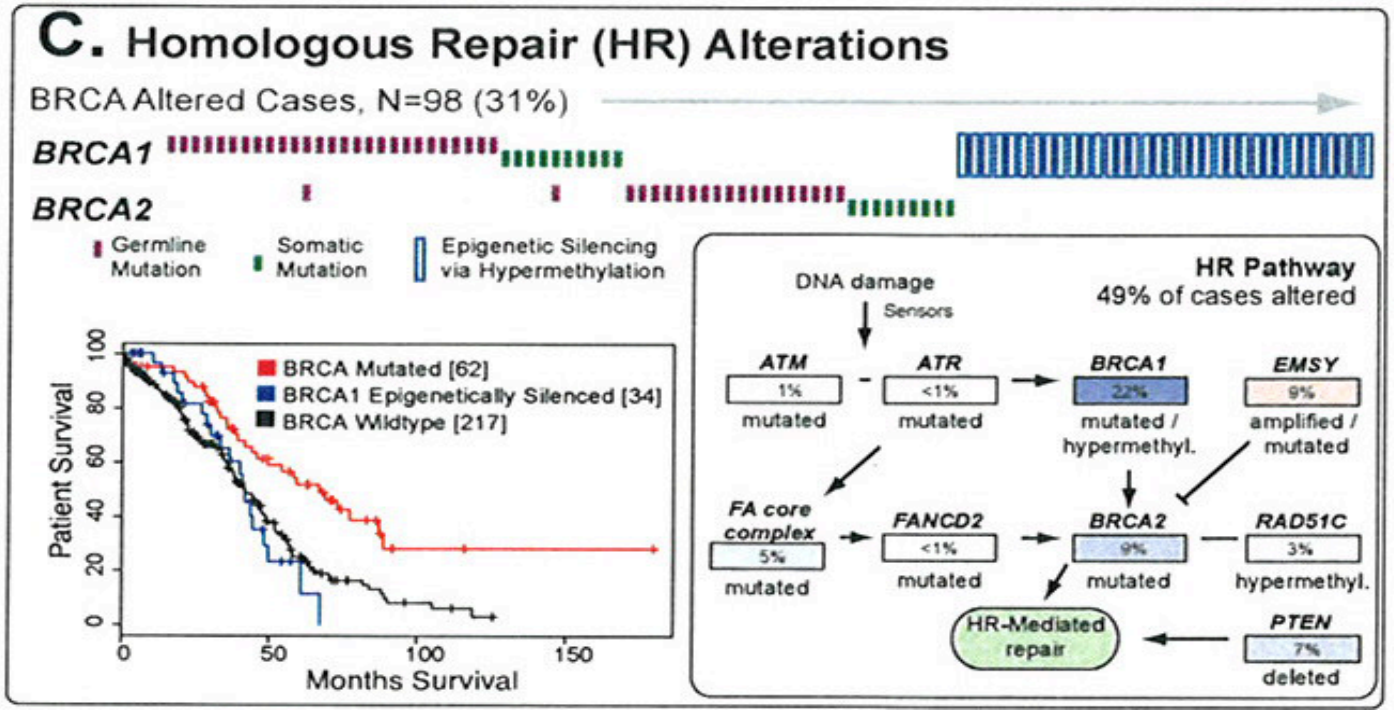
➔ OVCa is a disease of genomic instability driven by p53 mutation and defects in HR.

Patterns of somatic genomic alterations



- 68 amplified putative oncogenes in OVCa that are targets or putative targets of drugs or inhibitors in development

Pathway Analysis in Ovarian Cancer



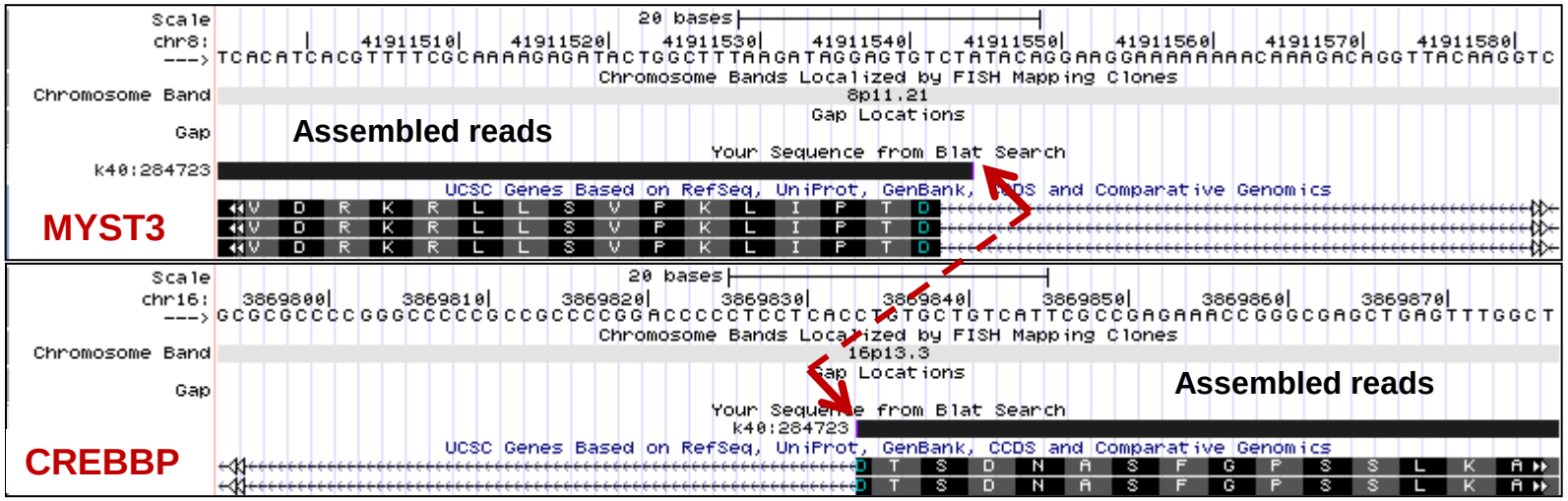
- Mutated BRCA1/2 have defective HR and are sensitive to PARP inhibitors
- HR defects occur in approximately half of serous OvCa
  - Core HR genes that are genomically altered
  - Mutation vs genomic amplification/deletion vs methylation

Fusion transcripts by RNA-seq in AML

THE CANCER GENOME ATLAS

- Identify by assembly and read pairs
  - AML1-ETO 5% of samples
  - PML-RARa 9% of samples
  - BCR-ABL 2% of samples
  - CBFB-MYH11 7% of samples
  - MLL fusions 5 to date
  - Other known 2 to date (CALM/AF10, MYST3/CREBBP)
  - Novel fusions 2 to date

Read assembly evidence for MYST3/ CREBBP fusion



Translocation in CRC by sequencing

THE CANCER GENOME ATLAS

Process:

~50bp Paired Ends Reads, Illumina HiSeq or GAII (4X Seq Coverage) → BWA alignment (.bam file) -l 40 -k 2 -n 3 → BreakDancer (structural variants)

Results:

Sequenced 10 pairs of Colorectal Cancer Pairs. We have analyzed 8 pairs so far. In red, these translocations are observed in multiple samples.

Translocations Detected by BreakDancer		Gene Name(s)
HFM1-SLCO5A1	*t(1;8)*	DNA helicase-SLCO5A1
HFM1-PPEF2	*t(1;4)*	DNA helicase-Protein Phosphatase
HFM1-SEC14L1	*t(1;17)*	DNA helicase-Sec 14 like
THEM2-SYN3	*t(6;22)*	Thioesterase-Synapsin
STMN3-KIAA1667	*t(20;22)*	Stathmin like-Herman Pudalski gene
MTMR2-LRCH1	*t(11;13)*	Myotubularin-Larch
PRDM9-PRDM7	*t(5;16)*	histone methyltransferase-histone methyltreansferase
THSD7B-C12orf32	*t(2;12)*	Thrombpspondin-Orf
NR5A2-KLHL29	*t(1;2)*	Nuclear Receptor-Kelch
WDR70-NXPH1	*t(5;7)*	WD Repeat-Neuroexophilin
C8orf37-CRTC1	*t(8;19)*	Orf-CREB related trx factor
SEMA5B-SPATS2	*t(3;12)*	Semaphorin-Spermatogenesis associated

Challenges ahead

- **Reference = Complete + Quality**
  - Quality: samples → biomolecules → data → analyses
  - Complete: Multi-dimensionality; global assays
  - Complete: sufficiently powered sample size
- **Analysis and Enablement**
  - Rapid data release
  - Analyses and Publication
  - Knowledge dissemination (Results, Tools)

Data analysis and dissemination

THE CANCER GENOME ATLAS

- ***Technical challenge***:
  - Cancer genomic data are noisy and complex, particularly challenging amidst rapid evolution in technological platforms
  - Better computational tools to make sense of the data
- ***Biological challenge***:
  - Cancer is biologically complex
  - Cancer gene functions are context specific

Genome Data Analysis Centers

Broad Institute, Cambridge, Mass.
Institute for Systems Biology, Seattle, Wash.
University of Texas/M.D. Anderson Cancer Center, Houston, Texas
Lawrence Berkeley National Laboratory, Berkeley, Calif.
Memorial Sloan-Kettering Cancer Center, New York, N.Y.
University of California at Santa Cruz, Calif.
University of North Carolina, Chapel Hill

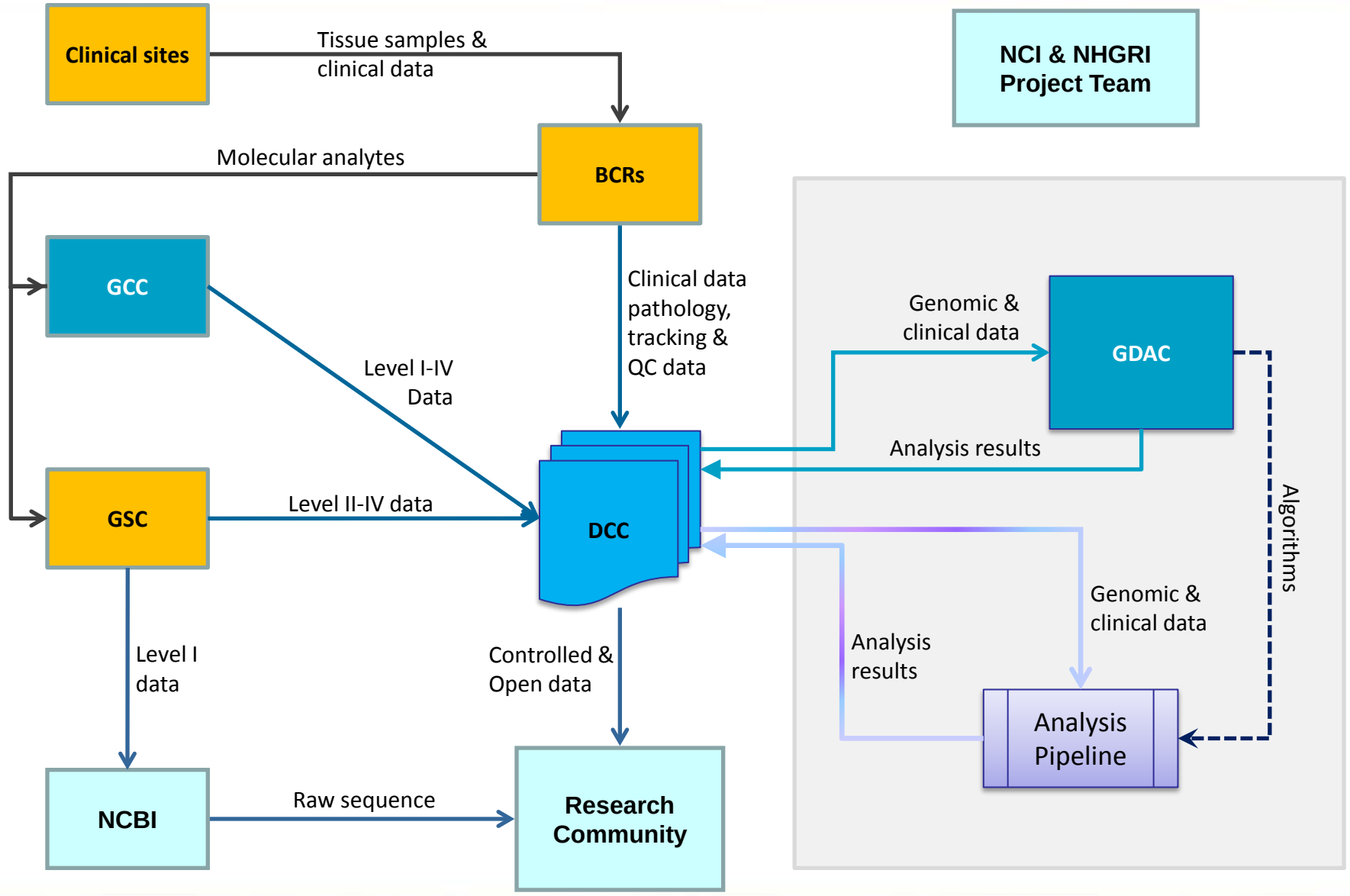
- ➔ Develop new computational tools for integrative cancer genome analyses
- ➔ Generate TCGA data analysis results in an "accessible" format for the cancer biology community
- ➔ Disseminate results rapidly

Analysis is a bottleneck

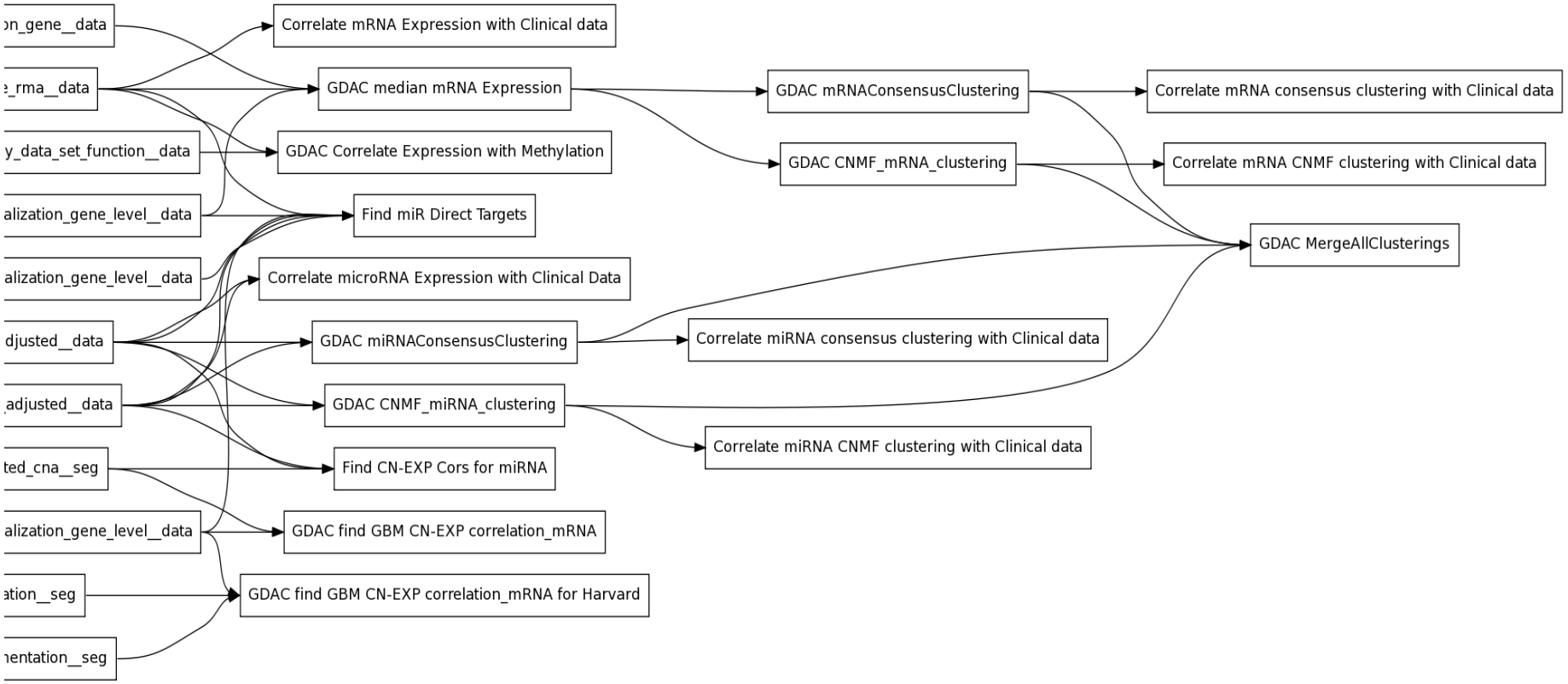
The Cancer Genome Atlas

Tumor Type	GCC assays	Whole Exomes	Whole Genomes
GBM	380	109 76 in progress	8 12 in progress
Ovarian	560	434 86 in progress	10 17 in progress
AML	162 39 in progress	15 135 in progress	26 29 in progress
Colon	103 41 in progress	52 51 in progress	0
Rectal	50 17 in progress	0 67 in progress	0
Breast ductal	0 233 in progress	0 186 in progress	0
Lung adeno	21 74 in progress	0 95 in progress	0
Lung scc	69 45 in progress	0 114 in progress	0
Endometrial	0 70 in progress	0 70 in progress	0
Renal	32	0 32 in progress	0
Gastric	0 82 in progress	0 82 in progress	0

TCGA Research Network



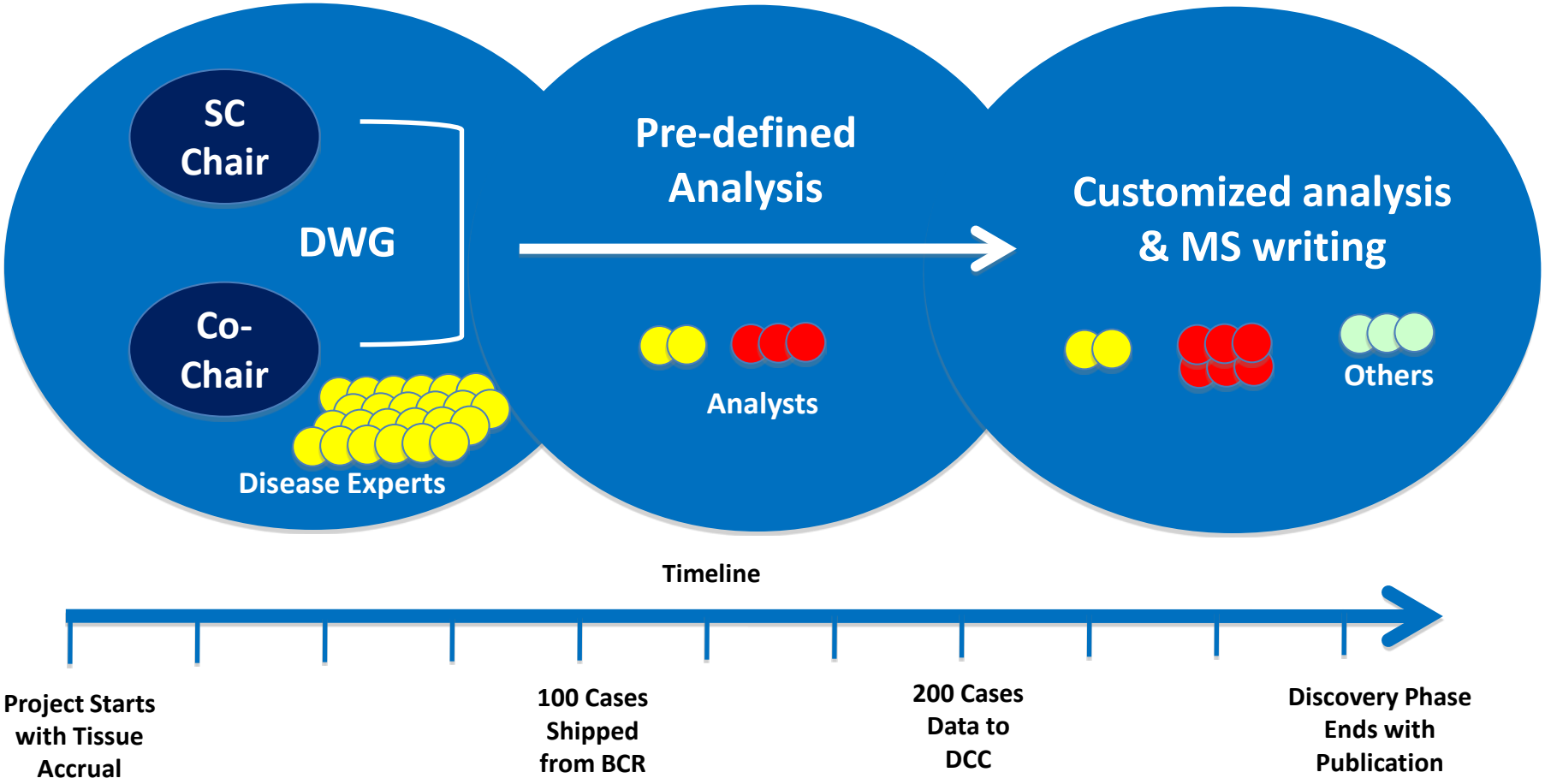
Example workflow of an analysis run



Input → **Automated Pre-defined Integrative Analyses** → **Output**

Mutation, copy number analysis; subclassification; pathway…

Streamlined Tumor Project Model



In the face of the evolving technologies…

THE CANCER GENOME ATLAS

- TCGA is generating new knowledge, enabling and impacting diverse research endeavors
- 'Genome Paradigm' brought to cancer
  - Completeness
  - Standardization
  - Open data release
- 'Field Enhancement' is evident
  - Methods improving
  - Costs driven down
  - Community engagement increasing
  - Log-changes being accepted and expected

Acknowledgement

THE CANCER GENOME ATLAS

Project Team
NCI and NHGRI

Genome Characterization Centers
SNP – Broad Institute
Genome Copy Number - Harvard
mRNA - Univ. North Carolina,
miRNA - Univ. British Columbia
Methylation - Univ. Southern Cali.
Adv. Genomics - Harvard, Baylor

Genome Sequencing Centers
Broad Institute
Washington University
Baylor College of Medicine

Data Coordinating Center

Genome Data Analysis Centers
Broad Institute, Institute for Systems Biology, MD Anderson Cancer Center, Lawrence Berkeley Nat'l Lab., Memorial Sloan Kettering Cancer Center, Univ. California, Santa Cruz, Univ. North Carolina

Public Data Portal
http://cancergenome.nih.gov/dataportal/