

# Mining Cancer Genomics Data

Aik Choon Tan, Ph.D.

Associate Professor of Bioinformatics

Division of Medical Oncology

Department of Medicine

[aikchoon.tan@ucdenver.edu](mailto:aikchoon.tan@ucdenver.edu)

11/6/2018

<http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640>

# Outline

---

- Motivation – Why Study Cancer Genomes
- Deep characterization of Cancer Genomes
- TCGA
- Pan-Cancer Genomes Analysis
- Tumor Heterogeneity
- Bioinformatics Tools

# Number of deaths for leading causes of death in US (CDC) (2011 data)

---

1. Heart disease: 596,339
2. Cancer: 575,313
3. Chronic lower respiratory diseases: 143,382
4. Stroke (cerebrovascular diseases): 128,931
5. Accidents (unintentional injuries): 122,777
6. Alzheimer's disease: 84,691
7. Diabetes: 73,282
8. Influenza and Pneumonia: 53,667
9. Nephritis, nephrotic syndrome, and nephrosis: 45,731
10. Intentional self-harm (suicide): 38,285

# Cancer: Number One Disease-Related Killer of Americans

---

Although extraordinary advances in cancer research have deepened our understanding of how cancer develops, grows, and threatens the lives of millions, it is projected that 580,350 Americans will die from one of the more than 200 types of cancer in 2013. Moreover, because more than 75 percent of cancer diagnoses occur in those aged 55 and older and this segment of the population is increasing in size, we face a future where the number of cancer-related deaths will increase dramatically. As a result, cancer is predicted to soon become the number one disease-related killer of Americans. This trend is being mirrored globally, and it is estimated that in 2030, more than 13 million people worldwide will lose their lives to cancer.

*(From: AACR Cancer Progress Report 2013)*

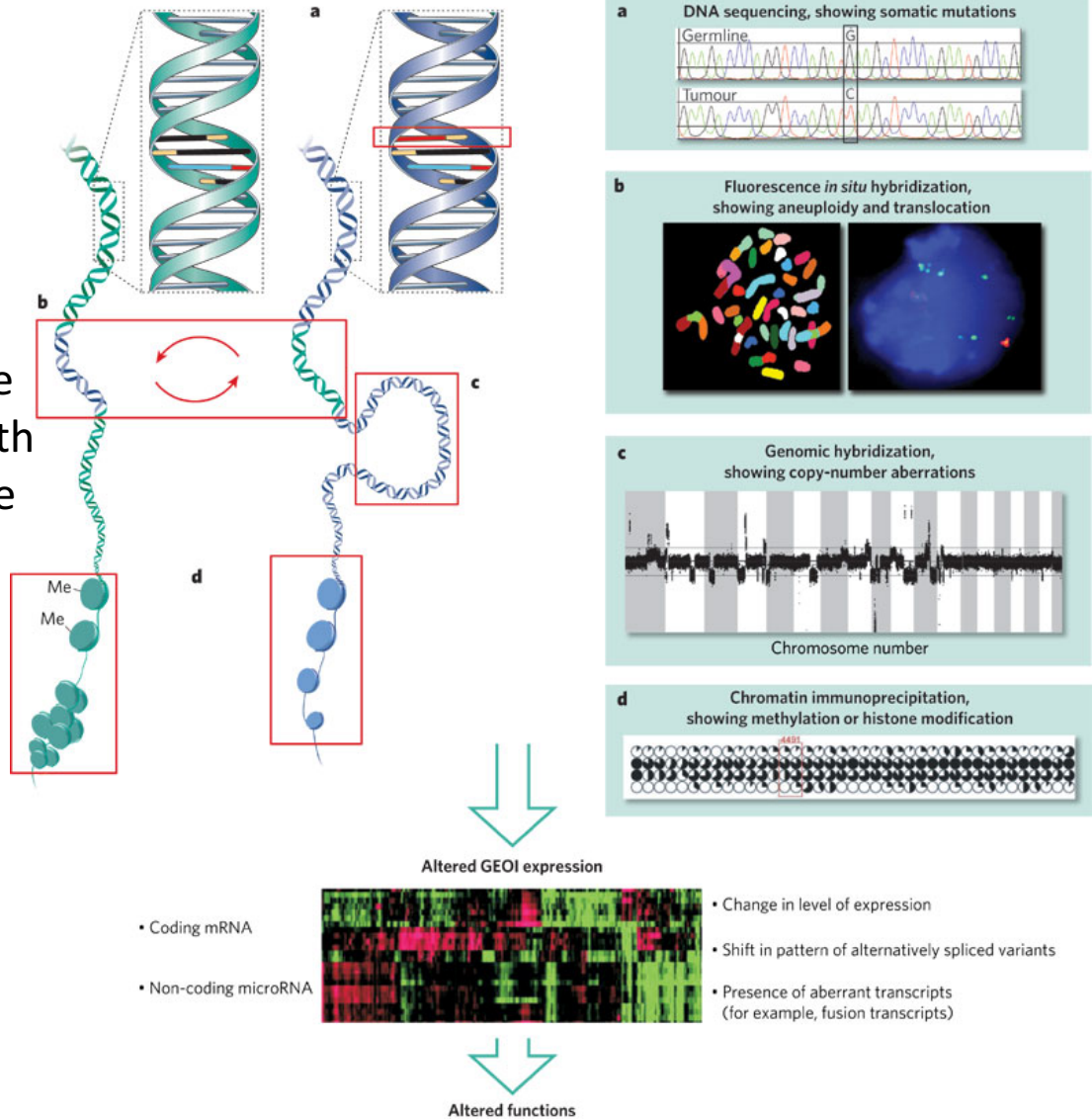
# Cancer

---

- Disease of Genome
  - Mutations in DNA (somatic mutations, copy number variations, translocations etc)
  - Alterations in Epigenetics (methylation patterns, epigenetics etc)
  - Alterations in RNA (mRNA expression, miRNA expression, other non-coding RNA expressions etc)
  - Alterations in Protein (protein expression, post-translational modification etc)
- Disease of Microenvironment
  - Signaling & Interactions from microenvironment

# Multiple Level of Aberrations in Cancer

Cancer is the phenotypic end point of numerous genomic and/or epigenomic alterations that have accumulated within cells, and of the interactions of such altered cells with the stromal components in a unique host microenvironment.



(Chin & Gray, Nature 2008)

# Drivers and Passengers

---

- Drivers

- Key genes when altered / mutated, can promote or “drive” tumorigenesis, and provide survival advantages to the cancer cell
- A typical tumor contains two to eight of these “driver gene” mutations. (Vogelstein et al Science 2013)
- Driver genes can be classified into 12 signaling pathways that regulate three core cellular processes: cell fate, cell survival, and genome maintenance. (Vogelstein et al Science 2013)

- Passengers

- Passenger mutations are “by-stander” alterations that happen to be altered in the primary cells but do not provide survival advantage of cancer.
- These mutations represent random somatic events (that is, changes that occurred before a clonal expansion and are simply carried along despite conferring no selective advantage).

- CHALLENGE: How to distinguish “drivers” from “passengers”

# Oncogenes addiction

---

- There are numerous examples which suggest that some cancers are not only initiated by a particular oncogene, but are also dependent upon that oncogene for tumor maintenance.
- Targeted cancer therapies have exploited this “*oncogene addiction*” concept; leading to several successful genotype-directed clinical applications of targeted therapies have been demonstrated
  - BCR-ABL in Chronic Myeloid Leukemia
  - EGFR in non small cell lung cancer (NSCLC)
  - EML-ALK4 in NSCLC
  - BRAF<sup>V600E</sup> in metastatic melanoma
- The clinical paradigm until now is based on a simple binary correlation between a mutated cancer gene and response to a given therapy.



# One Example: BRAF mutation in Melanoma

## Mutations of the *BRAF* gene in human cancer

(Nature 2002, 417:949-954)

Helen Davies<sup>1,2</sup>, Graham R. Bignell<sup>1,2</sup>, Charles Cox<sup>1,2</sup>, Philip Stephens<sup>1,2</sup>, Sarah Edkins<sup>1</sup>, Sheila Clegg<sup>1</sup>, Jon Teague<sup>1</sup>, Hayley Woffendin<sup>1</sup>, Mathew J. Garnett<sup>2</sup>, William Bottomley<sup>1</sup>, Neil Davis<sup>1</sup>, Ed Dicks<sup>1</sup>, Rebecca Ewing<sup>1</sup>, Yvonne Floyd<sup>1</sup>, Kristian Gray<sup>1</sup>, Sarah Hall<sup>1</sup>, Rachel Hawes<sup>1</sup>, Jaime Hughes<sup>1</sup>, Vivian Kosmidou<sup>1</sup>, Andrew Menzies<sup>1</sup>, Catherine Mould<sup>1</sup>, Adrian Parker<sup>1</sup>, Claire Stevens<sup>1</sup>, Stephen Watt<sup>1</sup>, Steven Hooper<sup>3</sup>, Rebecca Wilson<sup>3</sup>, Hiran Jayatilake<sup>4</sup>, Barry A. Gusterson<sup>5</sup>, Colin Cooper<sup>6</sup>, Janet Shipley<sup>6</sup>, Darren Hargrave<sup>7</sup>, Katherine Pritchard-Jones<sup>7</sup>, Norman Maitland<sup>8</sup>, Georgia Chenevix-Trench<sup>9</sup>, Gregory J. Riggins<sup>10</sup>, Darel D. Bigner<sup>10</sup>, Giuseppe Palmieri<sup>11</sup>, Antonio Cossu<sup>12</sup>, Adrienne Flanagan<sup>13</sup>, Andrew Nicholson<sup>14</sup>, Judy W. C. Ho<sup>15</sup>, Suet Y. Leung<sup>16</sup>, Siu T. Yuen<sup>16</sup>, Barbara L. Weber<sup>17</sup>, Hilliard F. Seigler<sup>18</sup>, Timothy L. Darrow<sup>18</sup>, Hugh Paterson<sup>3</sup>, Richard Marais<sup>3</sup>, Christopher J. Marshall<sup>3</sup>, Richard Wooster<sup>1,6</sup>, Michael R. Stratton<sup>1,4</sup> & P. Andrew Futreal<sup>1</sup>

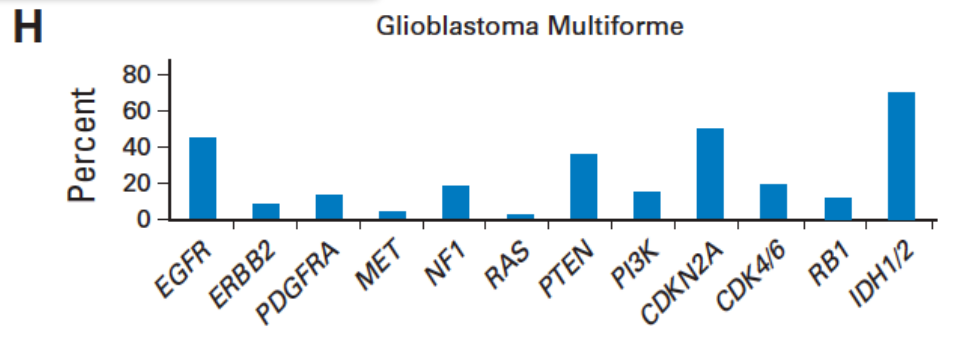
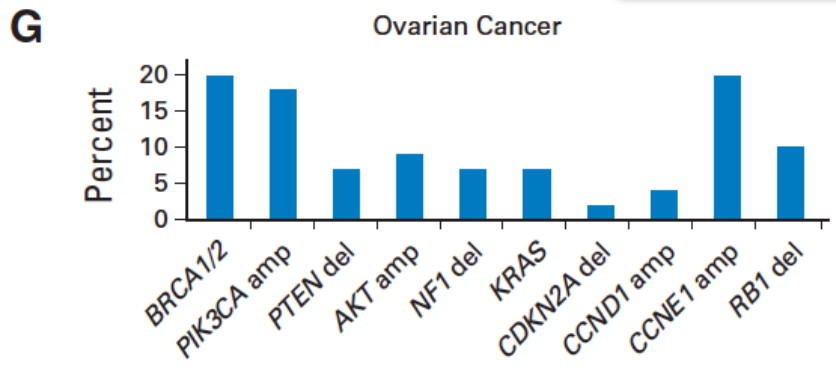
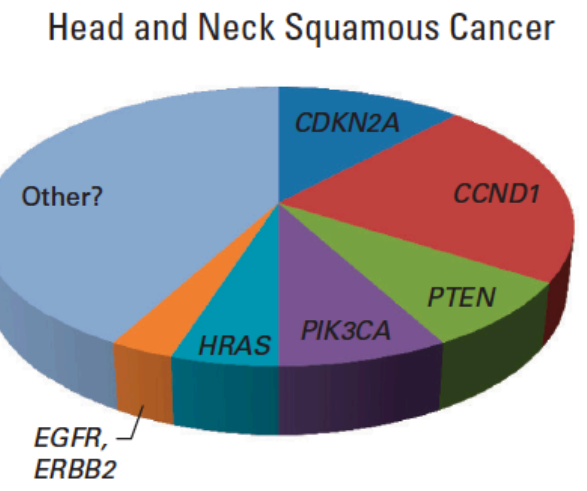
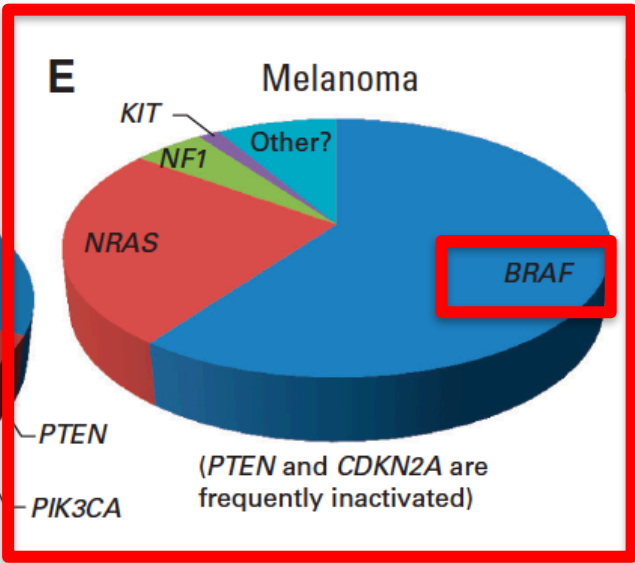
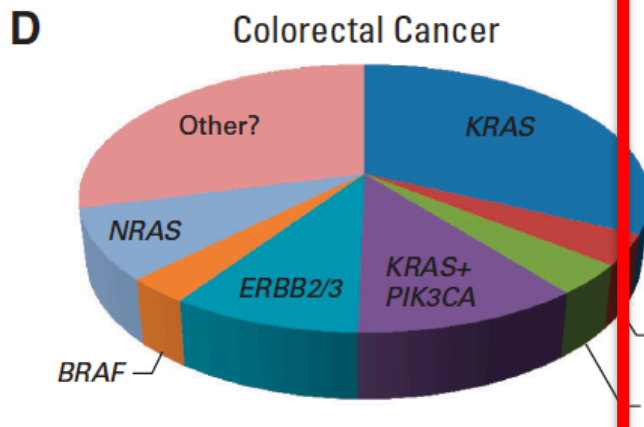
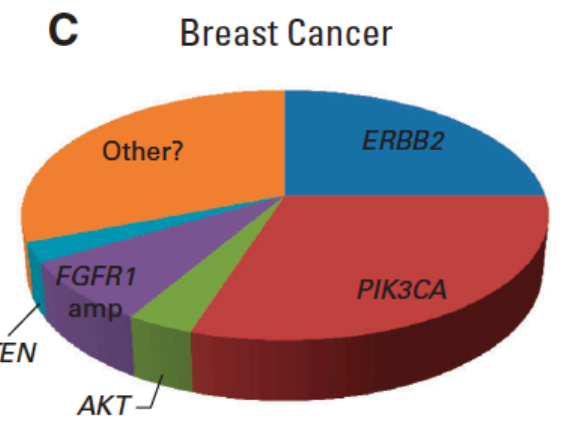
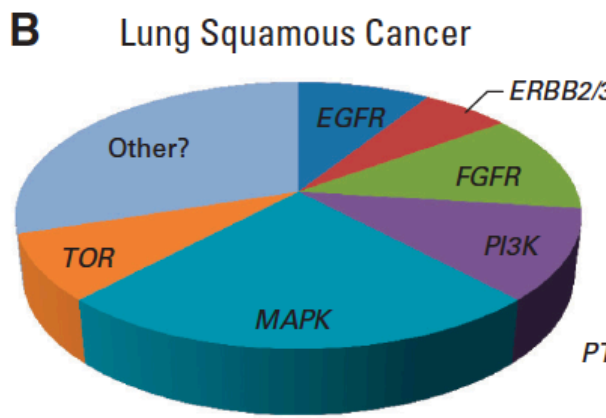
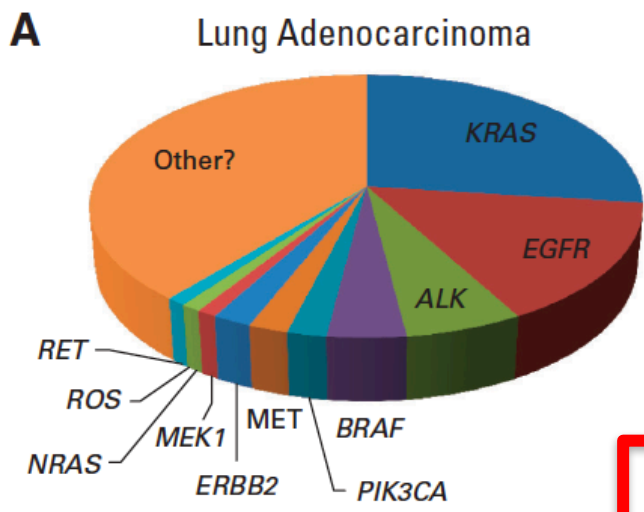
Table 1 *BRAF* mutations in human cancer

<i>BRAF</i> mutations		Cancer cell lines								Primary tumours						Total
Nucleotide	Amino acid	(1) Mel.	(2) Colo. ca.	(3) Glioma	(4) Lung ca.	(5) Sarcoma	(6) Breast	(7) Ovarian	(8) Other	(1) Mel. STC	(2) Mel.	(3) Colo. ca.	(4) Ovarian*	(5) Sarcoma	(6) Other†	
G1388A	G463E							1								1
G1388T	G463V		1													1
G1394C	G465A									1						1
G1394A	G465E										1					1
G1394T	G465V				1											1
G1403C	G468A				2											2
G1403A	G468E											1				1
G1753A	E585K												1			1
T1782G	F594L											1				1
G1783C	G595R		1													1
C1786G	L596V				1											1
T1787G	L596R												1			1
T1796A	V599E	19	5	4		5	1		1	11	5	2	3	1	0	57
TG1796-97AT	V599D	1														1
	Total	20	7	4	4	5	1	1	1	12	6	4	5	1	0	71
No. samples screened		34	40	38	131	59	45	26	172	15	9	33	35	182	104	923
Per cent		59%	18%	11%	3%	9%	2%	4%	0.6%	80%	67%	12%	14%	0.5%	0%	8%

Amino acid residues are grouped in blocks. Three further *BRAF* coding sequence variants were identified (G2041A R681Q in the HEC1A endometrial cancer cell line, T974C I325T in the ZR-75-30 breast cancer cell line, and C2180T A727V in the H33AJ-JA1 T-ALL cell line). These were not present in 341 control DNAs. Lane numbers (in parentheses) are provided for convenience. Mel., melanoma; Colo. ca., colorectal cancer; Mel. STC, melanoma short-term culture.

\*Four out of ten LMP (low malignant potential); 1 out of 25 malignant epithelial.

†Glioma ( $n = 15$ ), breast cancer ( $n = 33$ ), prostate cancer ( $n = 23$ ), HNSCC (head and neck squamous cell carcinoma) ( $n = 19$ ), lung cancer ( $n = 14$ ).



# MELANOMA PROGRESSION

Mutations

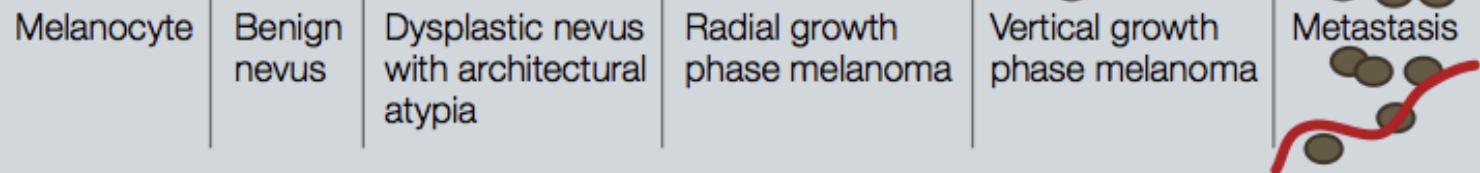
**BRAF**, NRAS,  
GNAQ, KIT\*

Inactivation of *CDKN2A*, *PTEN*;  
Activation of *MITF*, *CDK4*, *CCND1*

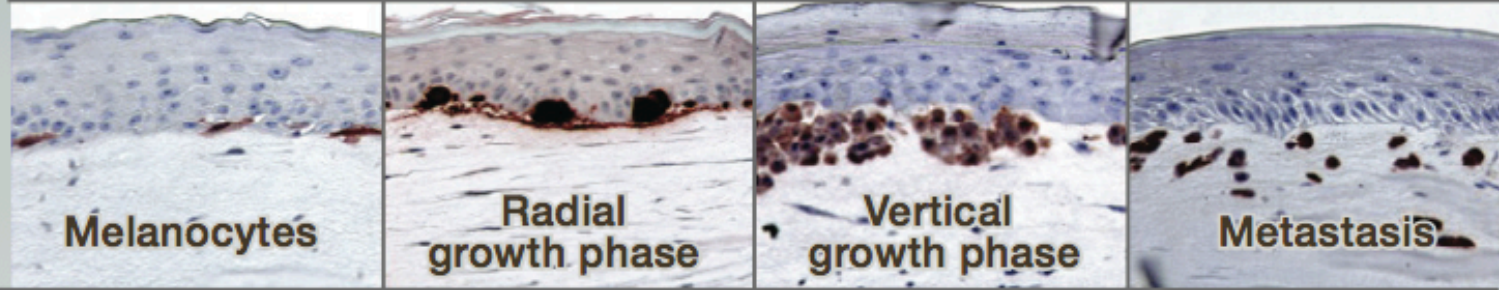
\*Mutually exclusive in most cases.

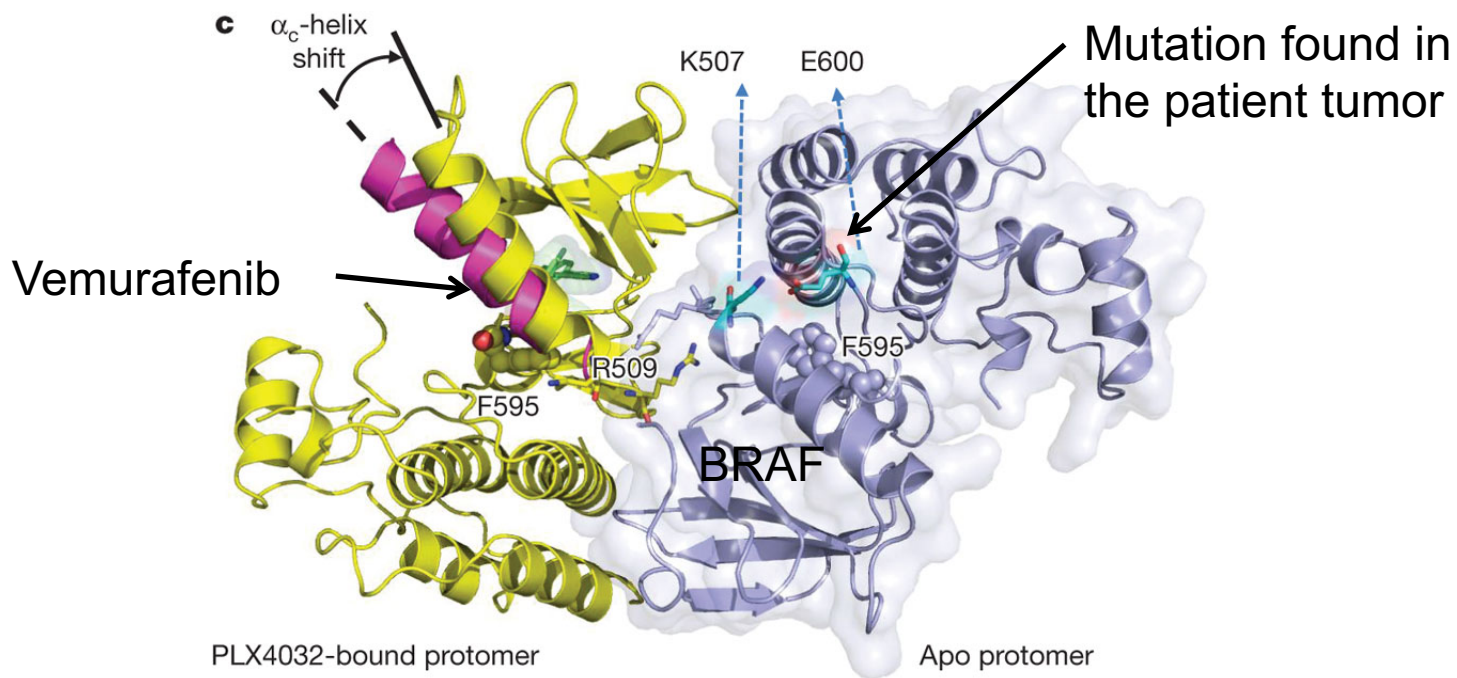
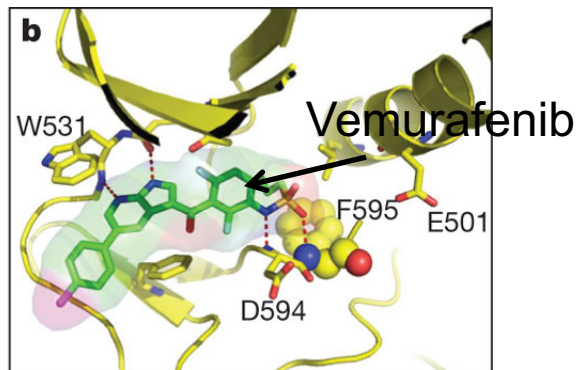
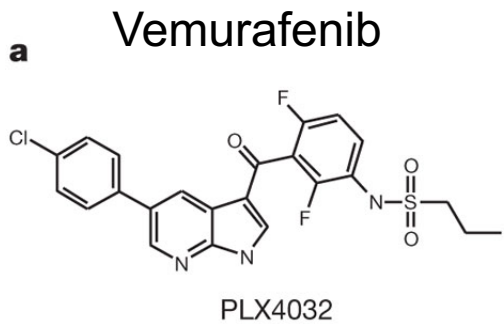
Epidermis  
Basement  
membrane

Dermis

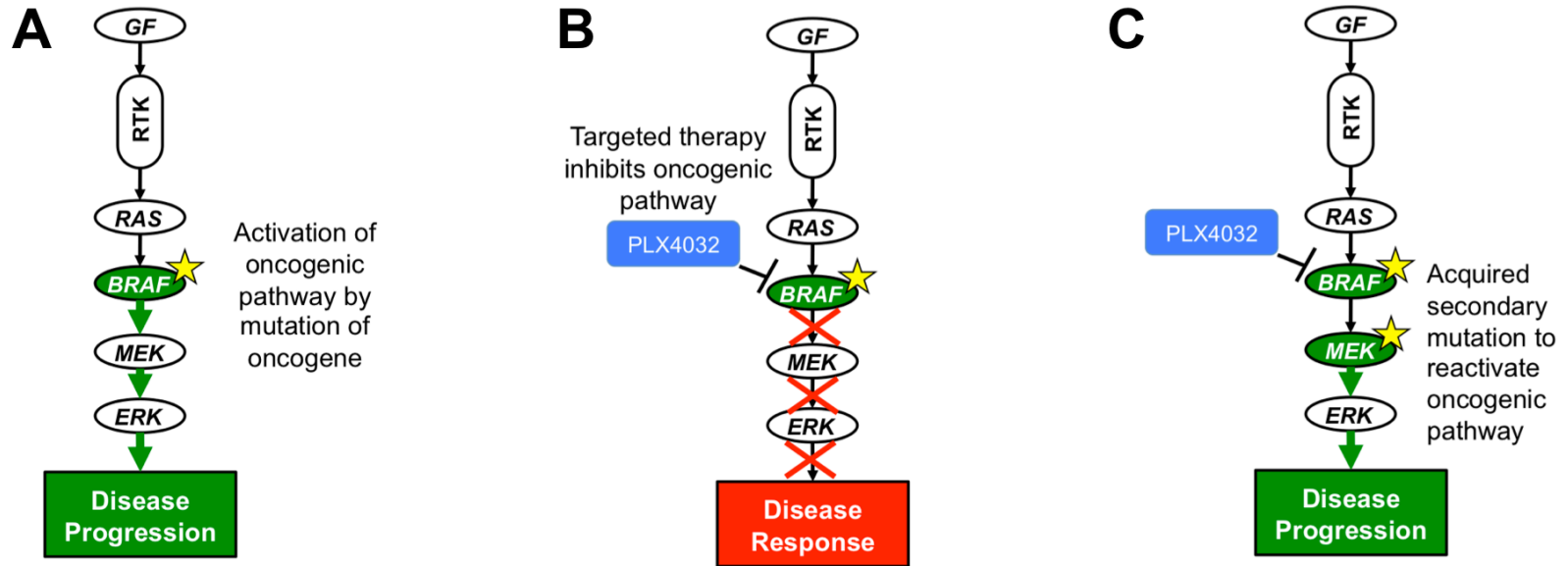


Skin reconstruct  
cross-section





# Targeting mutated genes with Targeted Therapy



**Pre-treatment**

(38 yo melanoma patient with BRAF mut)



**Post-treatment**

15 weeks on PLX4032 (BRAFi)



**Relapse (acquired resistance)**

23 weeks on PLX4032 (BRAFi)

(Patient Image From: *Wagle et al JCO 2011*)

ORIGINAL ARTICLE

# Combined BRAF and MEK Inhibition in Melanoma with BRAF V600 Mutations

Keith T. Flaherty, M.D., Jeffery R. Infante, M.D., Adil Daud, M.D., Rene Gonzalez, M.D., Richard F. Kefford, M.D., Ph.D., Jeffrey Sosman, M.D., Omid Hamid, M.D., Lynn Schuchter, M.D., Jonathan Cebon, M.D., Ph.D., Nageatte Ibrahim, M.D., Ragini Kudchadkar, M.D., Howard A. Burris III, M.D., Gerald Falchook, M.D., Alain Algazi, M.D., Karl Lewis, M.D., Georgina V. Long, M.D., Ph.D., Igor Puzanov, M.D., M.S.C.I., Peter Lebowitz, M.D., Ph.D., Ajay Singh, M.D., Shonda Little, M.P.H., Peng Sun, Ph.D., Alicia Allred, Ph.D., Daniele Ouellet, Ph.D., Kevin B. Kim, M.D., Kiran Patel, M.D., M.B.A., and Jeffrey Weber, M.D., Ph.D.

ABSTRACT

**BACKGROUND**

Resistance to therapy with BRAF kinase inhibitors is associated with reactivation of the mitogen-activated protein kinase (MAPK) pathway. To address this problem, we conducted a phase 1 and 2 trial of combined treatment with dabrafenib, a selective BRAF inhibitor, and trametinib, a selective MAPK kinase (MEK) inhibitor.

**METHODS**

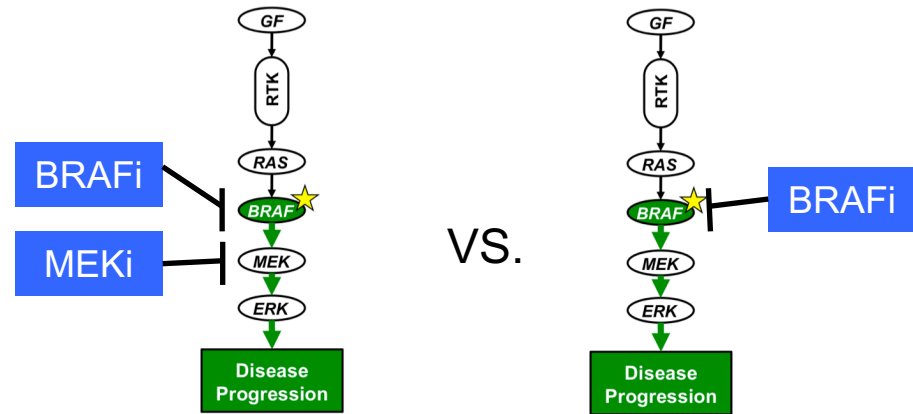
In this open-label study involving 247 patients with metastatic melanoma and BRAF V600 mutations, we evaluated the pharmacokinetic activity and safety of oral dabrafenib (75 or 150 mg twice daily) and trametinib (1, 1.5, or 2 mg daily) in 85 patients and then randomly assigned 162 patients to receive combination therapy with dabrafenib (150 mg) plus trametinib (1 or 2 mg) or dabrafenib monotherapy. The primary end points were the incidence of cutaneous squamous-cell carcinoma, survival free of melanoma progression, and response. Secondary end points were overall survival and pharmacokinetic activity.

**RESULTS**

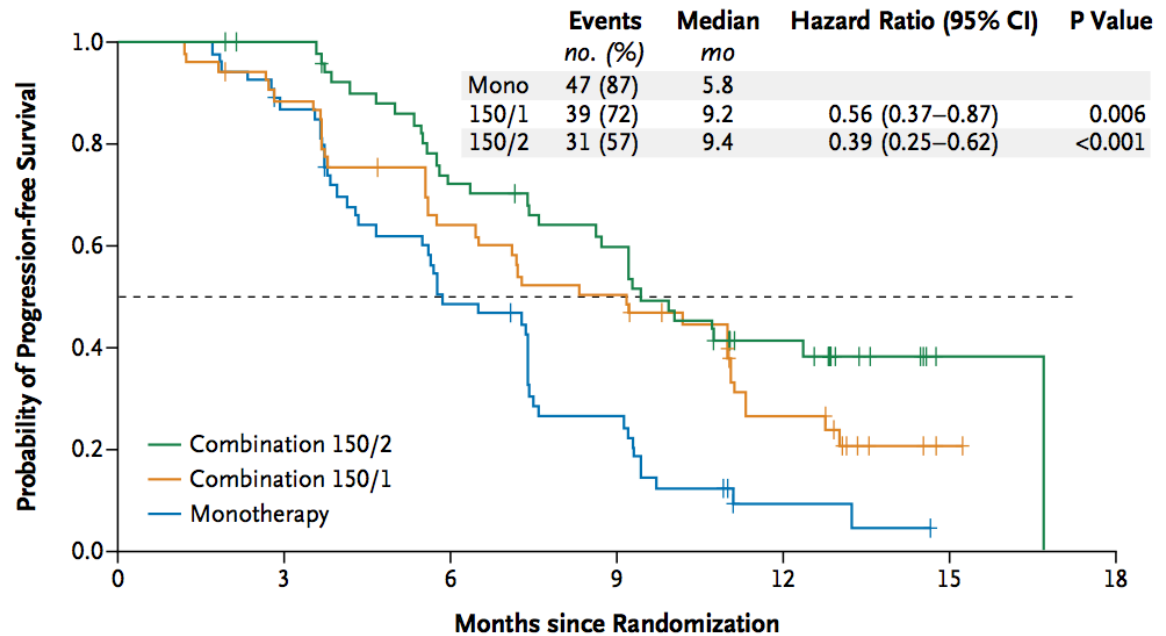
Dose-limiting toxic effects were infrequently observed in patients receiving combination therapy with 150 mg of dabrafenib and 2 mg of trametinib (combination 150/2). Cutaneous squamous-cell carcinoma was seen in 7% of patients receiving combination 150/2 and in 19% receiving monotherapy (P=0.09), whereas pyrexia was more common in the combination 150/2 group than in the monotherapy group (71% vs. 26%). Median progression-free survival in the combination 150/2 group was 9.4 months, as compared with 5.8 months in the monotherapy group (hazard ratio for progression or death, 0.39; 95% confidence interval, 0.25 to 0.62; P<0.001). The rate of complete or partial response with combination 150/2 therapy was 76%, as compared with 54% with monotherapy (P=0.03).

**CONCLUSIONS**

Dabrafenib and trametinib were safely combined at full monotherapy doses. The rate of pyrexia was increased with combination therapy, whereas the rate of proliferative skin lesions was nonsignificantly reduced. Progression-free survival was significantly improved. (Funded by GlaxoSmithKline; ClinicalTrials.gov number, NCT01072175.)



## Progression Free Survival



&gt;&gt; COSMIC Database

ARTICLES

# Patterns of somatic mutation in human cancer genomes

Christopher Greenman<sup>1</sup>, Philip Stephens<sup>1</sup>, Raffaella Smith<sup>1</sup>, Gillian L. Dalglish<sup>1</sup>, Christopher Hunter<sup>1</sup>, Graham Bignell<sup>1</sup>, Helen Davies<sup>1</sup>, Jon Teague<sup>1</sup>, Adam Butler<sup>1</sup>, Claire Stevens<sup>1</sup>, Sarah Edkins<sup>1</sup>, Sarah O'Meara<sup>1</sup>, Imre Vastrik<sup>2</sup>, Esther E. Schmidt<sup>2</sup>, Tim Avis<sup>1</sup>, Syd Barthorpe<sup>1</sup>, Gurpreet Bhamra<sup>1</sup>, Gemma Buck<sup>1</sup>, Bhudipa Choudhury<sup>1</sup>, Jody Clements<sup>1</sup>, Jennifer Cole<sup>1</sup>, Ed Dicks<sup>1</sup>, Simon Forbes<sup>1</sup>, Kris Gray<sup>1</sup>, Kelly Halliday<sup>1</sup>, Rachel Harrison<sup>1</sup>, Katy Hills<sup>1</sup>, Jon Hinton<sup>1</sup>, Andy Jenkinson<sup>1</sup>, David Jones<sup>1</sup>, Andy Menzies<sup>1</sup>, Tatiana Mironenko<sup>1</sup>, Janet Perry<sup>1</sup>, Keiran Raine<sup>1</sup>, Dave Richardson<sup>1</sup>, Rebecca Shepherd<sup>1</sup>, Alexandra Small<sup>1</sup>, Calli Tofts<sup>1</sup>, Jennifer Varian<sup>1</sup>, Tony Webb<sup>1</sup>, Sofie West<sup>1</sup>, Sara Widaa<sup>1</sup>, Andy Yates<sup>1</sup>, Daniel P. Cahill<sup>3</sup>, David N. Louis<sup>3</sup>, Peter Goldstraw<sup>4</sup>, Andrew G. Nicholson<sup>4</sup>, Francis Brasseur<sup>5</sup>, Leendert Looijenga<sup>6</sup>, Barbara L. Weber<sup>7</sup>, Yoke-Eng Chiew<sup>8</sup>, Anna deFazio<sup>8</sup>, Mel F. Greaves<sup>9</sup>, Anthony R. Green<sup>10</sup>, Peter Campbell<sup>1</sup>, Ewan Birney<sup>2</sup>, Douglas F. Easton<sup>11</sup>, Georgia Chenevix-Trench<sup>12</sup>, Min-Han Tan<sup>13</sup>, Sok Kean Khoo<sup>13</sup>, Bin Tean Teh<sup>13</sup>, Siu Tsan Yuen<sup>14</sup>, Suet Yi Leung<sup>14</sup>, Richard Wooster<sup>1</sup>, P. Andrew Futreal<sup>1</sup> & Michael R. Stratton<sup>1,9</sup>

Cancers arise owing to mutations in a subset of genes that confer growth advantage. The availability of the human genome sequence led us to propose that systematic resequencing of cancer genomes for mutations would lead to the discovery of many additional cancer genes. Here we report more than 1,000 somatic mutations found in 274 megabases (Mb) of DNA corresponding to the coding exons of 518 protein kinase genes in 210 diverse human cancers. There was substantial variation in the number and pattern of mutations in individual cancers reflecting different exposures, DNA repair defects and cellular origins. Most somatic mutations are likely to be 'passengers' that do not contribute to oncogenesis. However, there was evidence for 'driver' mutations contributing to the development of the cancers studied in approximately 120 genes. Systematic sequencing of cancer genomes therefore reveals the evolutionary diversity of cancers and implicates a larger repertoire of cancer genes than previously anticipated.

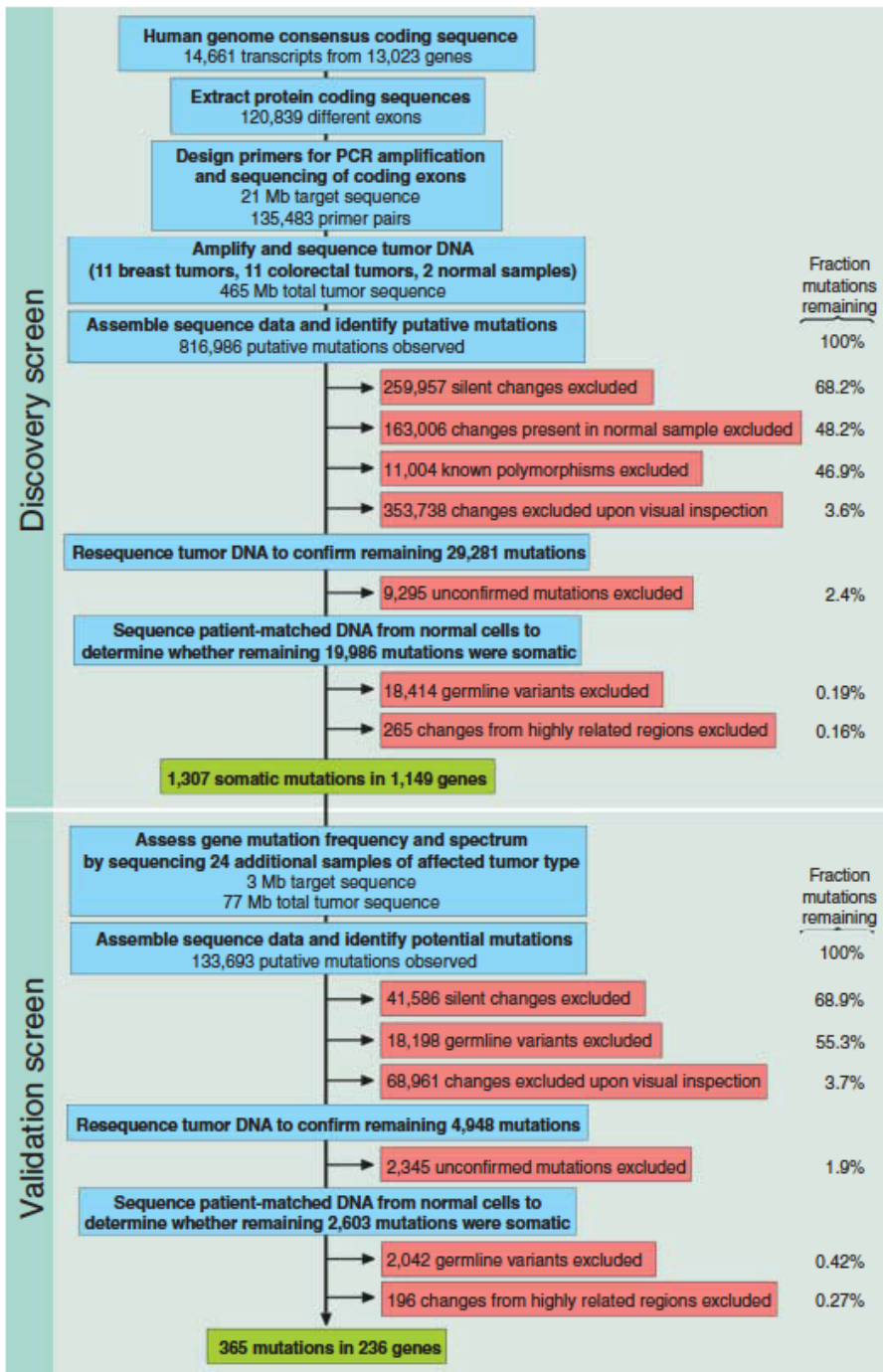
## RESEARCH ARTICLE

# The Consensus Coding Sequences of Human Breast and Colorectal Cancers

Tobias Sjöblom,<sup>1\*</sup> Siân Jones,<sup>1\*</sup> Laura D. Wood,<sup>1\*</sup> D. Williams Parsons,<sup>1\*</sup> Jimmy Lin,<sup>1</sup> Thomas D. Barber,<sup>1†</sup> Diana Mandelker,<sup>1</sup> Rebecca J. Leary,<sup>1</sup> Janine Ptak,<sup>1</sup> Natalie Silliman,<sup>1</sup> Steve Szabo,<sup>1</sup> Phillip Buckhaults,<sup>2</sup> Christopher Farrell,<sup>2</sup> Paul Meeh,<sup>2</sup> Sanford D. Markowitz,<sup>3</sup> Joseph Willis,<sup>4</sup> Dawn Dawson,<sup>4</sup> James K. V. Willson,<sup>5</sup> Adi F. Gazdar,<sup>6</sup> James Hartigan,<sup>7</sup> Leo Wu,<sup>8</sup> Changsheng Liu,<sup>8</sup> Giovanni Parmigiani,<sup>9</sup> Ben Ho Park,<sup>10</sup> Kurtis E. Bachman,<sup>11</sup> Nickolas Papadopoulos,<sup>1</sup> Bert Vogelstein,<sup>1‡</sup> Kenneth W. Kinzler,<sup>1‡</sup> Victor E. Velculescu<sup>1‡</sup>

The elucidation of the human genome sequence has made it possible to identify genetic alterations in cancers in unprecedented detail. To begin a systematic analysis of such alterations, we determined the sequence of well-annotated human protein-coding genes in two common tumor types. Analysis of 13,023 genes in 11 breast and 11 colorectal cancers revealed that individual tumors accumulate an average of ~90 mutant genes but that only a subset of these contribute to the neoplastic process. Using stringent criteria to delineate this subset, we identified 189 genes (average of 11 per tumor) that were mutated at significant frequency. The vast majority of these genes were not known to be genetically altered in tumors and are predicted to affect a wide range of cellular functions, including transcription, adhesion, and invasion. These data define the genetic landscape of two human cancer types, provide new targets for diagnostic and therapeutic intervention, and open fertile avenues for basic research in tumor biology.





# Initial Large-Scale Cancer Genome Sequencing Efforts

## RESEARCH ARTICLES

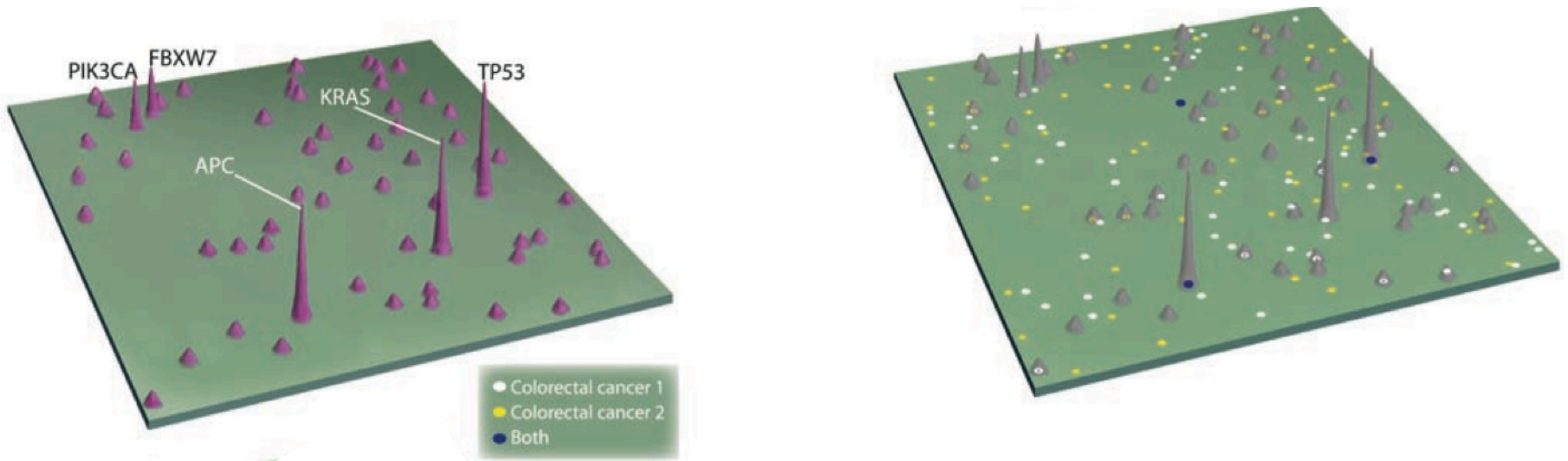
### The Genomic Landscapes of Human Breast and Colorectal Cancers

Laura D. Wood,<sup>1\*</sup> D. Williams Parsons,<sup>1\*</sup> Siân Jones,<sup>1\*</sup> Jimmy Lin,<sup>1\*</sup> Tobias Sjöblom,<sup>1\*†</sup> Rebecca J. Leary,<sup>1</sup> Dong Shen,<sup>1</sup> Simina M. Boca,<sup>1,2</sup> Thomas Barber,<sup>1‡</sup> Janine Ptak,<sup>1</sup> Natalie Silliman,<sup>1</sup> Steve Szabo,<sup>1</sup> Zoltan Dezso,<sup>3</sup> Vadim Ustyansky,<sup>3</sup> Tatiana Nikolskaya,<sup>3,4</sup> Yuri Nikolsky,<sup>3</sup> Rachel Karchin,<sup>5</sup> Paul A. Wilson,<sup>5</sup> Joshua S. Kaminker,<sup>6</sup> Zemin Zhang,<sup>6</sup> Randal Croshaw,<sup>7</sup> Joseph Willis,<sup>8</sup> Dawn Dawson,<sup>8</sup> Michail Shipitsin,<sup>9</sup> James K. V. Willson,<sup>10</sup> Saraswati Sukumar,<sup>11</sup> Kornelia Polyak,<sup>9</sup> Ben Ho Park,<sup>11</sup> Charit L. Pethiyagoda,<sup>12</sup> P. V. Krishna Pant,<sup>12</sup> Dennis G. Ballinger,<sup>12</sup> Andrew B. Sparks,<sup>12§</sup> James Hartigan,<sup>13</sup> Douglas R. Smith,<sup>13</sup> Erick Suh,<sup>13</sup> Nickolas Papadopoulos,<sup>1</sup> Phillip Buckhaults,<sup>7</sup> Sanford D. Markowitz,<sup>14</sup> Giovanni Parmigiani,<sup>1||</sup> Kenneth W. Kinzler,<sup>1||</sup> Victor E. Velculescu,<sup>1||</sup> Bert Vogelstein<sup>1||</sup>

(Science 2007)

**Human cancer is caused by the accumulation of mutations in oncogenes and tumor suppressor genes. To catalog the genetic changes that occur during tumorigenesis, we isolated DNA from 11 breast and 11 colorectal tumors and determined the sequences of the genes in the Reference Sequence database in these samples. Based on analysis of exons representing 20,857 transcripts from 18,191 genes, we conclude that the genomic landscapes of breast and colorectal cancers are composed of a handful of commonly mutated gene “mountains” and a much larger number of gene “hills” that are mutated at low frequency. We describe statistical and bioinformatic tools that may help identify mutations with a role in tumorigenesis. These results have implications for understanding the nature and heterogeneity of human cancers and for using personal genomics for tumor diagnosis and therapy.**

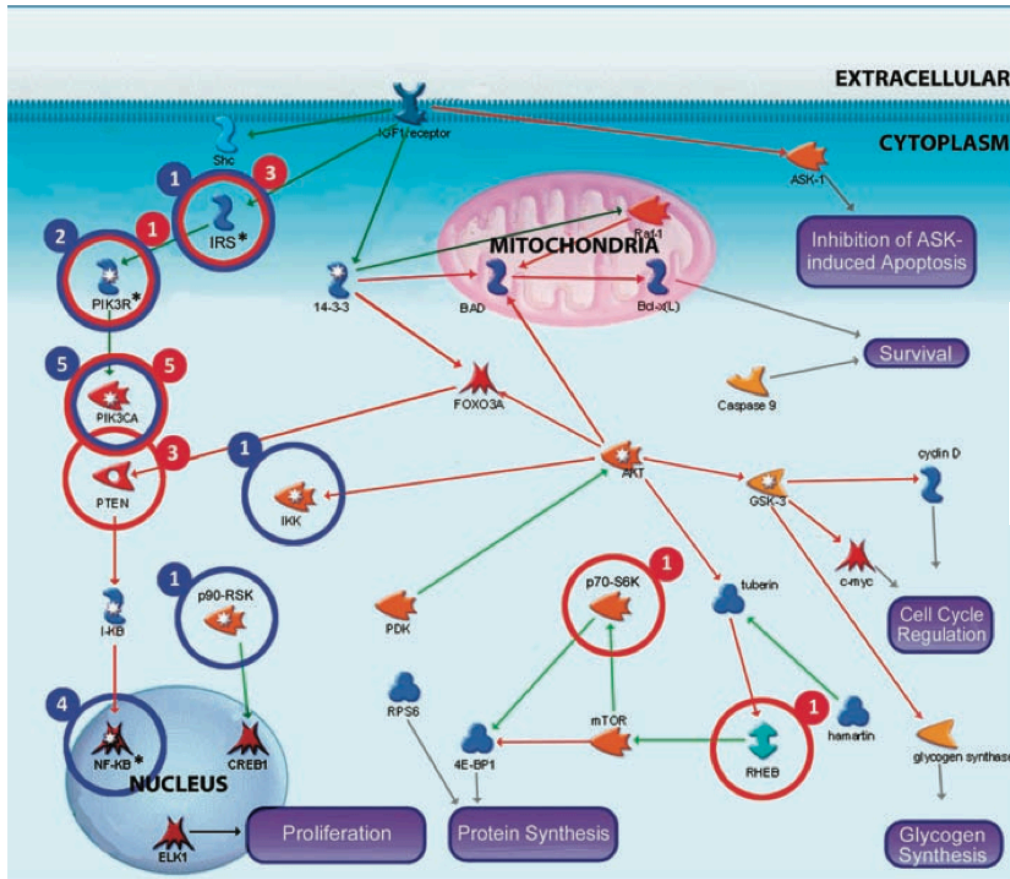
# Initial Large-Scale Cancer Genome Sequencing Efforts



**Table 1.** Summary of somatic mutations. UTR, untranslated region. ND, not determined because synonymous mutations were not evaluated in the RefSeq genes analyzed in (5).

Tumor type	Screen	Gene set	Mutated genes	Coding changes						Noncoding changes	Total mutations
				Missense	Nonsense	Insertion	Deletion	Duplication	Synonymous	Splice site or UTR	
Colorectal cancers	Discovery	This study	325	237	14	0	8	0	93	12	364
		All RefSeq	848	722	48	4	27	18	ND	30	942
	Validation	This study	88	81	9	1	2	2	30	6	131
		All RefSeq	183	197	34	4	14	5	ND	15	299
Breast cancers	Discovery	This study	460	304	26	2	28	1	131	14	506
		All RefSeq	1137	909	64	5	78	3	ND	53	1243
	Validation	This study	62	52	3	0	3	0	19	2	79
		All RefSeq	167	153	11	2	15	2	ND	7	209

# Pathway-centric analysis of Cancer Genome



Regardless of whether this pathway-centric interpretation is correct, it is clear that the “easy” part of future cancer genome research will be the identification of genetic alterations. The vast

**Fig. 2.** PI3K pathway mutations in breast and colorectal cancers. The identities and relationships of genes that function in PI3K signaling are indicated. Circled genes have somatic mutations in colorectal (red) and breast (blue) cancers. The number of tumors with somatic mutations in each mutated protein is indicated by the number adjacent to the circle. Asterisks indicate proteins with mutated isoforms that may play similar roles in the cell. These include insulin receptor substrates IRS2 and IRS4; phosphatidylinositol 3-kinase regulatory subunits PIK3R1, PIK3R4, and PIK3R5; and NF-κB regulators NFKB1, NFKBIA, and NFKBIE.

# The Cancer Genome Atlas Project

---

- The Cancer Genome Atlas (TCGA) began as a three-year pilot in 2006 with an investment of \$50 million each from the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI).
- Pilot Project: Comprehensive Characterized of three tumor types: Glioblastoma (GBM), Ovarian and lung cancers (> 200 samples)
- 2009 expand to 20 common tumor types

## BOX 1 TCGA: MISSION AND STRATEGY

Important information about the biological relevance of the molecular changes in cancer can be obtained through combined analysis of multiple different types of data.

For that reason, TCGA's principal aims are to generate, quality control, merge, analyze and interpret molecular profiles at the DNA, RNA, protein and epigenetic levels for hundreds of clinical tumors representing various tumor types and their subtypes. Cases that meet quality assurance specifications are characterized using technologies that assess the sequence of the exome, copy number variation (measured by SNP arrays), DNA methylation, mRNA expression and sequence, microRNA expression and transcript splice variation. Additional platforms applied to a subset of the tumors, including whole-genome sequencing and RPPAs, provide additional layers of data to complement the core genomic data sets and clinical data. By the end of 2015, the TCGA Research Network plans to have achieved the ambitious goal of analyzing the genomic, epigenomic and gene expression profiles of more than 10,000 specimens from more than 25 different tumor types.

TCGA has other, complementary aims as well: to promote the development and application of new technologies, to detect cancer-specific molecular alterations, to make data and results freely available to the scientific community, to develop tools and standard operating procedures that can serve other large-scale profiling projects and to build cadres of individuals (including experimentalists, computational biologists, statistical analysts, computer scientists and administrative staff) with the expertise to carry out such large-scale, team science projects. As of 24 July 2013, TCGA had mapped molecular patterns across 7,992 total cases representing 27 tumor types. The data, along with tools for exploring them, are publicly available at <http://www.cancergenome.nih.gov/>. Eight 'marker papers' (comprehensive initial publications on each of the tumor types) have been published so far<sup>8-12,14,27</sup>.

# The Cancer Genome Atlas (TCGA)

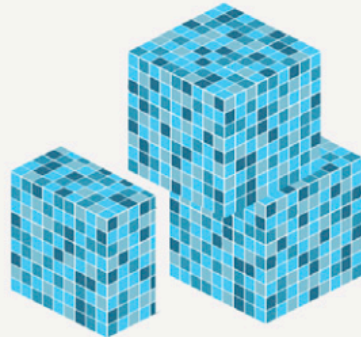
## THE CANCER GENOME ATLAS (TCGA) BY THE NUMBERS

TCGA produced over

**2.5**

PETABYTES

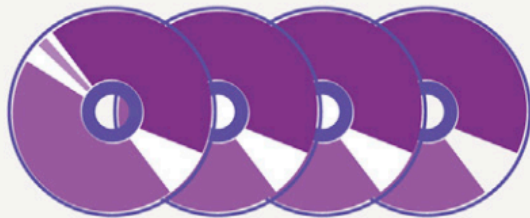
of data



To put this into perspective, **1 petabyte** of data is equal to

**212,000**

DVDs



TCGA data describes



**33**

DIFFERENT  
TUMOR TYPES

...including

**10**

RARE  
CANCERS

...based on paired tumor and normal tissue sets collected from



**11,000**

PATIENTS

...using

**7**

DIFFERENT  
DATA TYPES



Cancer type	Prevalence	TCGA cases assessed	Key findings
Breast lobular carcinoma	3,327,552	203	FOXA1 elevated in lobular carcinoma, GATA3 in ductal carcinoma; lobular enriched for PTEN loss and Akt activation
Breast ductal carcinoma		784	Four distinct subtypes: basal, Her2, luminal A, and luminal B; most common driver mutations: TP53, PIK3CA, GATA3; basal subtype similar to serous ovarian cancer
Prostate cancer	3,085,209	333	Highly heterogeneous with 26% driven by unknown alterations; ETS gene fusions or mutations in SPOR, FOXA1, or IDH1 define seven subtypes; actionable lesions in PI3K, MAPK, and DNA repair pathways
Colorectal adenocarcinoma	1,317,247	276	Colon and rectal cancers have similar genomic profiles; hypermutated subtype associated with favorable prognosis; new potential drivers: ARID1A, SOX9, FAM123B/WTX
Cutaneous melanoma	1,169,351	331	Established four subtypes: BRAF mutant, RAS mutant, NF1 mutant, and triple wild-type based on driver mutations; higher levels of immune lymphocyte infiltration correlated with better survival
Thyroid carcinoma	726,646	496	Majority driven by RAS or BRAFV600E mutations
Endometrial carcinoma	710,228	373	Classified endometrial cancers into four categories: POLE ultramutated, MSI (microsatellite instability) hypermutated, copy-number low, copy-number high
Uterine carcinosarcoma		57	Strong and varied degree of epithelial-mesenchymal transition; TP53 mutations in 91% of samples; PI3K alterations in half
Invasive urothelial bladder carcinoma	696,440	131	Increased risk associated with smoking; frequently mutated; TP53 inactivated in 76% of tumors, ERBB2 (HER2), genes in the RTK/RAS pathways altered in 44%
Lung adenocarcinoma	527,228	230	High mutation burden; 76% have activation of receptor tyrosine kinase pathways
Lung squamous cell carcinoma		178	High average number of mutations and copy-number aberrations; almost all have mutation in TP53; many have inactivating mutations in HLA-A that may aid immune evasion
Clear cell renal cell carcinoma	483,225	446	Commonly mutated genes: VHL, SED2, and the PI3K/AKT/mTOR pathway; metabolic shift similar to the "Warburg effect" correlates with a poor prognosis
Kidney papillary carcinoma		161	81% of type 1 tumors had MET alteration; type 2 tumors were heterogeneous, with alterations to CDKN2A, SETD2, TFE3, or increased expression of NRF2-ARE pathway; loss of CDKN2A expression and CpG island methylation phenotype associated with poor outcome
Chromophobe renal cell carcinoma		66	Extremely low mutation burden; metabolic shift distinct from the "Warburg effect" shift in clear cell carcinoma; TP53 and PTEN were frequently mutated; TERT gene promoter was frequently altered
Cervical cancer	256,078	228	Identification of HPV-negative, endometrial-like cancers with mutations in KRAS, ARID1A, and PTEN; amplification of CD274 and PDCD1LG2; frequent alterations in MED1, ERBB3, CASP8, HLA-A, and TGFBR2 and fusions involving lncRNA BCAR4; nearly three-quarters had alterations in either or both of the PI3K/MAPK and TGF-beta pathways
Testicular germ cell cancer	251,194	150	
Ovarian serous adenocarcinoma	222,060	489	Mutations: TP53 (in 96%), BRCA1 and BRCA2 (in 21%) and are associated with more favorable outcomes
Glioblastoma multiforme	162341	206	GBM subtypes Classical, Mesenchymal, and Proneural are defined by EGFR, NF1, and PDGFRA/IDH1 mutations, respectively; over 40% have mutations in chromatin-modifiers; frequently mutated: TP53, PIK3R1, PIK3CA, IDH1, PTEN, RB1, LZTR1
Lower-grade glioma		293	Defined three subtypes correlating with patient outcomes: IDH1 mutant with 1p/19q deletion, IDH mutant without 1p/19q deletion, and IDH wild-type
Stomach adenocarcinoma	95,764	295	Identified four subtypes characterized by EBV infection, microsatellite instability, genomic stability, and chromosomal instability
Liver hepatocellular carcinoma	66,771	363	TERT promoter mutations in 44%; TP53 commonly mutated or under-expressed; CTNBB1 significantly mutated; many tumors had high levels of lymphocyte infiltration or overexpressed immune checkpoint genes
Cholangiocarcinoma		38	Low expression of CDKN2, BAP1, and ARID1 genes and overexpression of FGFR2 and IDH1/2 genes; four subtypes defined
Pancreatic ductal adenocarcinoma	64,668	150	KRAS mutations present in 93% of tumors; mutations in RREB1 or other members of RAS-MAPK signaling pathway
Esophageal carcinoma	45,547	164	Squamous cell carcinomas had frequent amplifications of CCND1, SOX2, and TP63; adenocarcinomas had frequent amplifications in ERBB2, VEGFA, GATA4, and GATA6
Acute myeloid leukemia		200	Low mutation burden—only 13 coding mutations on average per tumor; classified driver events into nine categories including transcription factor fusions, histone modifier mutations, and spliceosome mutations
Head and neck squamous cell carcinoma		279	HPV-positive associated with shortened or deleted TRAF3, HPV-negative characterized by co-amplification of 11q13 and 11q22, smoking-related characterized by TP53 mutations, CDKN2A inactivation, CNVs
Sarcoma		206	TP53, ATRX, and RB1 are recurrently mutated across all types; synovial sarcomas expressed fusions in SSB1 or SSB2 and TERT; JUN amplification associates with worse survival in dedifferentiated liposarcoma
Paraganglioma and pheochromocytoma		173	Four distinct subtypes: Wnt-altered, cortical admixture, pseudohypoxia, and kinase signaling; MAML3 fusion gene and CSDE1 somatic mutation define and drive the poor prognosis Wnt-altered subtype
Thymoma		124	
Adrenocortical carcinoma		91	Overexpression of IGF2, mutations in TP53, PRKAR1A and other genes, and copy-number alterations were common; hypoploidy followed by whole-genome doubling may be a driving mechanism
Mesothelioma		87	
Uveal melanoma		80	Complex mutation in BAP1 gene; identified distinct subdivisions of disomy 3 (D3) and monosomy 3 (M3) subtypes; in M3, mutually exclusive EIF1AX and SRSF2/SF3B1 mutations have distinct methylation profiles and prognoses



# The Cancer Genome Atlas (TCGA)

## RESULTS & FINDINGS



### MOLECULAR BASIS OF CANCER

Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.



### TUMOR SUBTYPES

Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.\*



### THERAPEUTIC TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

# The Cancer Genome Atlas (TCGA)

## THE TEAM



# 20

## COLLABORATING INSTITUTIONS

across the United States  
and Canada

## WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



\*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

# Components of the TCGA Research Network

---

- **Biospecimen Core Resource (BCR)** – Tissue samples are carefully cataloged, processed, checked for quality and stored, complete with important medical information about the patient.
- **Genome Characterization Centers (GCCs)** – Several technologies will be used to analyze genomic changes involved in cancer. The genomic changes that are identified will be further studied by the Genome Sequencing Centers.
- **Genome Sequencing Centers (GSCs)** – High-throughput Genome Sequencing Centers will identify the changes in DNA sequences that are associated with specific types of cancer.
- **Proteome Characterization Centers (PCCs)** – The centers, a component of NCI's Clinical Proteomic Tumor Analysis Consortium, will ascertain and analyze the total proteomic content of a subset of TCGA samples.
- **Data Coordinating Center (DCC)** – The information that is generated by TCGA will be centrally managed at the DCC and entered into the TCGA Data Portal and Cancer Genomics Hub as it becomes available. Centralization of data facilitates data transfer between the network and the research community, and makes data analysis more efficient. The DCC manages the TCGA Data Portal.
- **Cancer Genomics Hub (CGHub)** – Lower level sequence data will be deposited into a secure repository. This database stores cancer genome sequences and alignments.
- **Genome Data Analysis Centers (GDACs)** – Immense amounts of data from array and second-generation sequencing technologies must be integrated across thousands of samples. These centers will provide novel informatics tools to the entire research community to facilitate broader use of TCGA data.

# CGC Centers and Data Types

---

Center	Data Type
BCGAC	Illumina miRNA-seq
BROAD	SNP 6.0 (copy number)
HMS Harvard	Illumina DNA-seq
JHU/USC	Methylation
MD Anderson	RPPA
UNC	Agilent Microarray (gene exp)
	Illumina RNA-seq

# Comprehensive Molecular Characterization of Colorectal Cancer

ARTICLE

doi:10.1038/nature11252

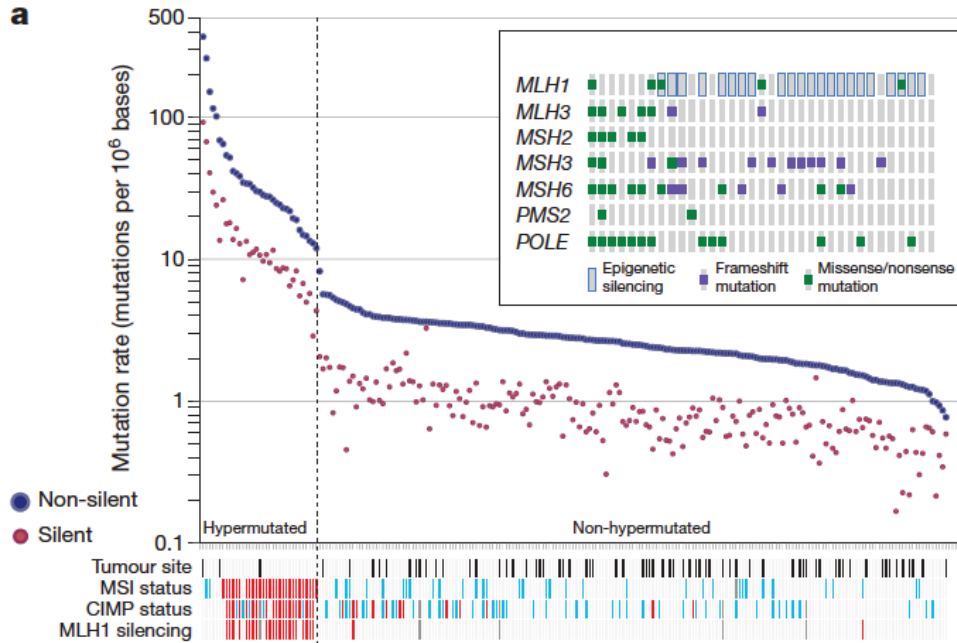
## Comprehensive molecular characterization of human colon and rectal cancer

The Cancer Genome Atlas Network\*

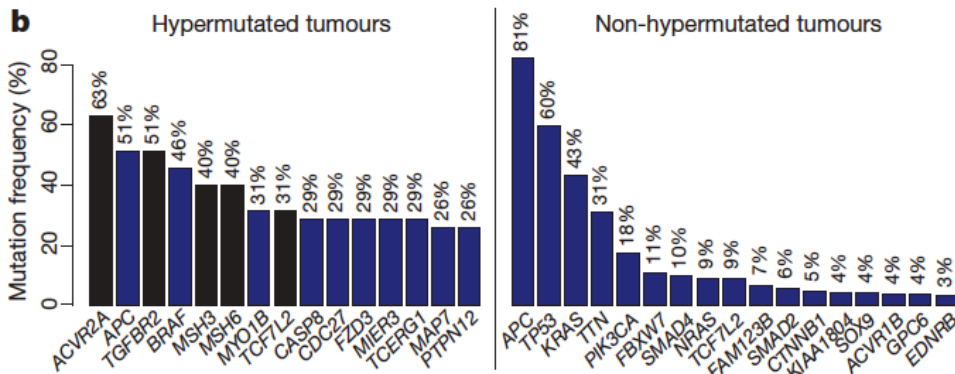
To characterize somatic alterations in colorectal carcinoma, we conducted a genome-scale analysis of 276 samples, analysing exome sequence, DNA copy number, promoter methylation and messenger RNA and microRNA expression. A subset of these samples (97) underwent low-depth-of-coverage whole-genome sequencing. In total, 16% of colorectal carcinomas were found to be hypermutated: three-quarters of these had the expected high microsatellite instability, usually with hypermethylation and *MLH1* silencing, and one-quarter had somatic mismatch-repair gene and polymerase  $\epsilon$  (*POLE*) mutations. Excluding the hypermutated cancers, colon and rectum cancers were found to have considerably similar patterns of genomic alteration. Twenty-four genes were significantly mutated, and in addition to the expected *APC*, *TP53*, *SMAD4*, *PIK3CA* and *KRAS* mutations, we found frequent mutations in *ARID1A*, *SOX9* and *FAM123B*. Recurrent copy-number alterations include potentially drug-targetable amplifications of *ERBB2* and newly discovered amplification of *IGF2*. Recurrent chromosomal translocations include the fusion of *NAV2* and WNT pathway member *TCF7L1*. Integrative analyses suggest new markers for aggressive colorectal carcinoma and an important role for *MYC*-directed transcriptional activation and repression.

This file contains the legends for Supplementary Tables 1-12, Supplementary Tables 1-9 and Supplementary Data files 1-2, Supplementary Methods, which include 17 Figures and 2 Tables (see Contents for details) and Supplementary Figures 1-9.

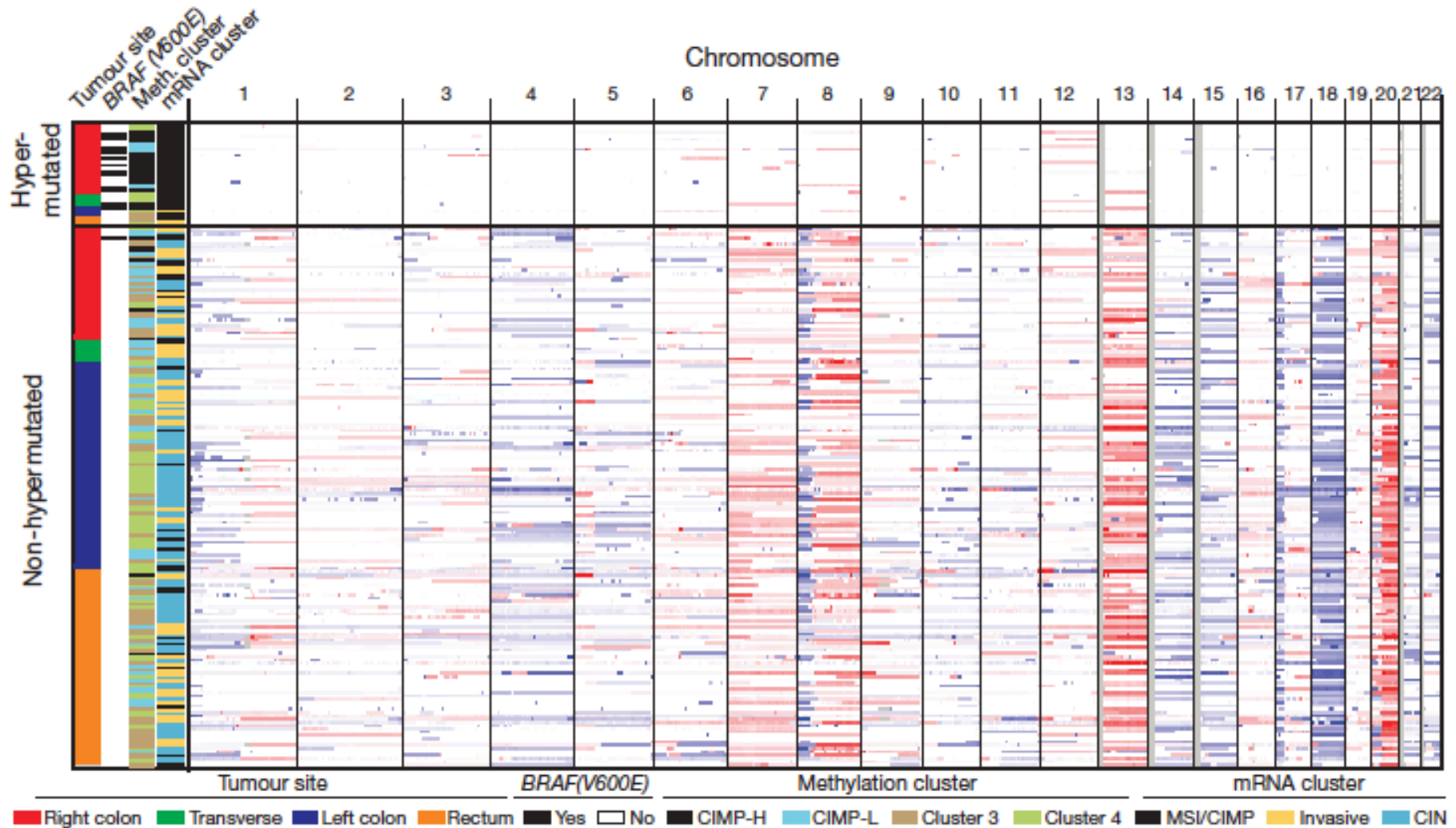
# Mutation Frequency



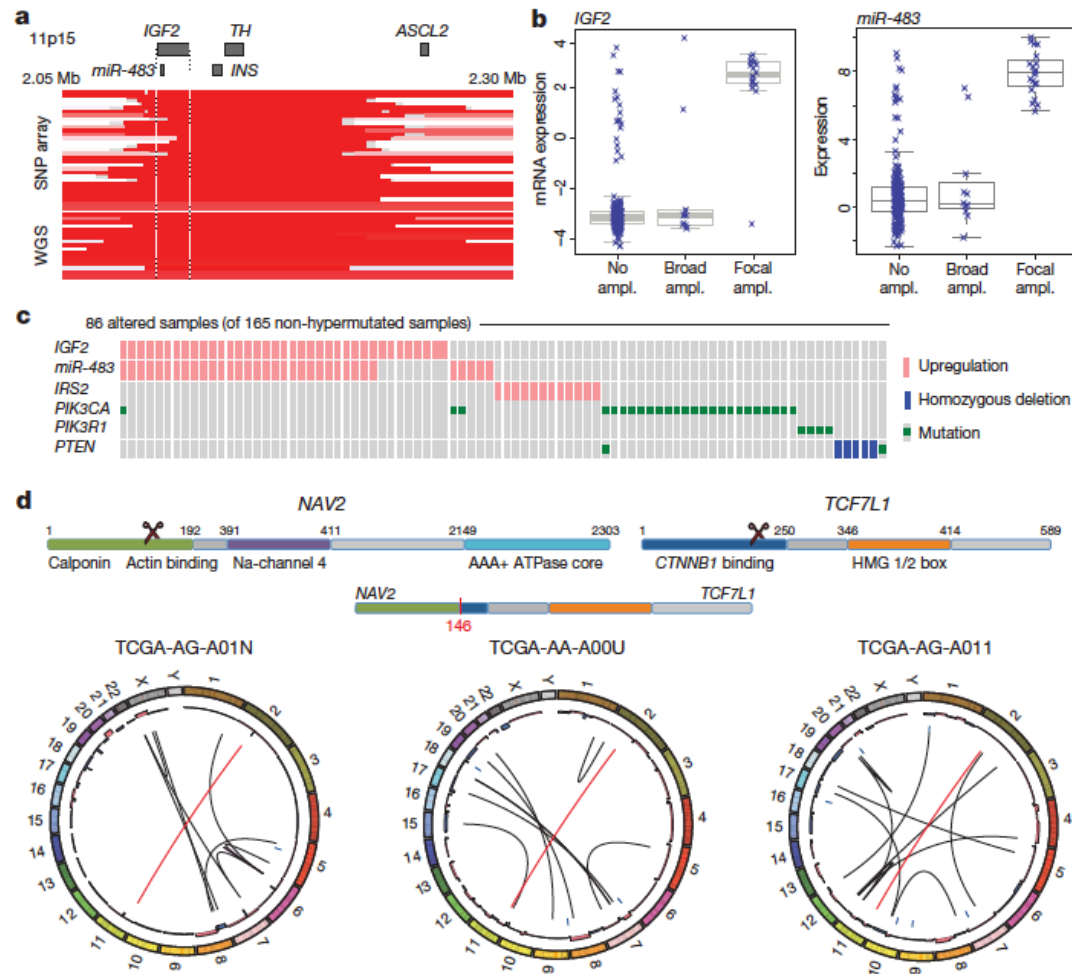
**Figure 1 | Mutation frequencies in human CRC.** **a**, Mutation frequencies in each of the tumour samples from 224 patients. Note a clear separation of hypermutated and non-hypermutated samples. Red, MSI high, CIMP high or MLH1 silenced; light blue, MSI low, or CIMP low; black, rectum; white, colon; grey, no data. Inset, mutations in mismatch-repair genes and *POLE* among the hypermutated samples. The order of the samples is the same as in the main graph. **b**, Significantly mutated genes in hypermutated and non-hypermutated tumours. Blue bars represent genes identified by the MutSig algorithm and black bars represent genes identified by manual examination of sequence data.



# Integrative Analysis



# Novel Insights

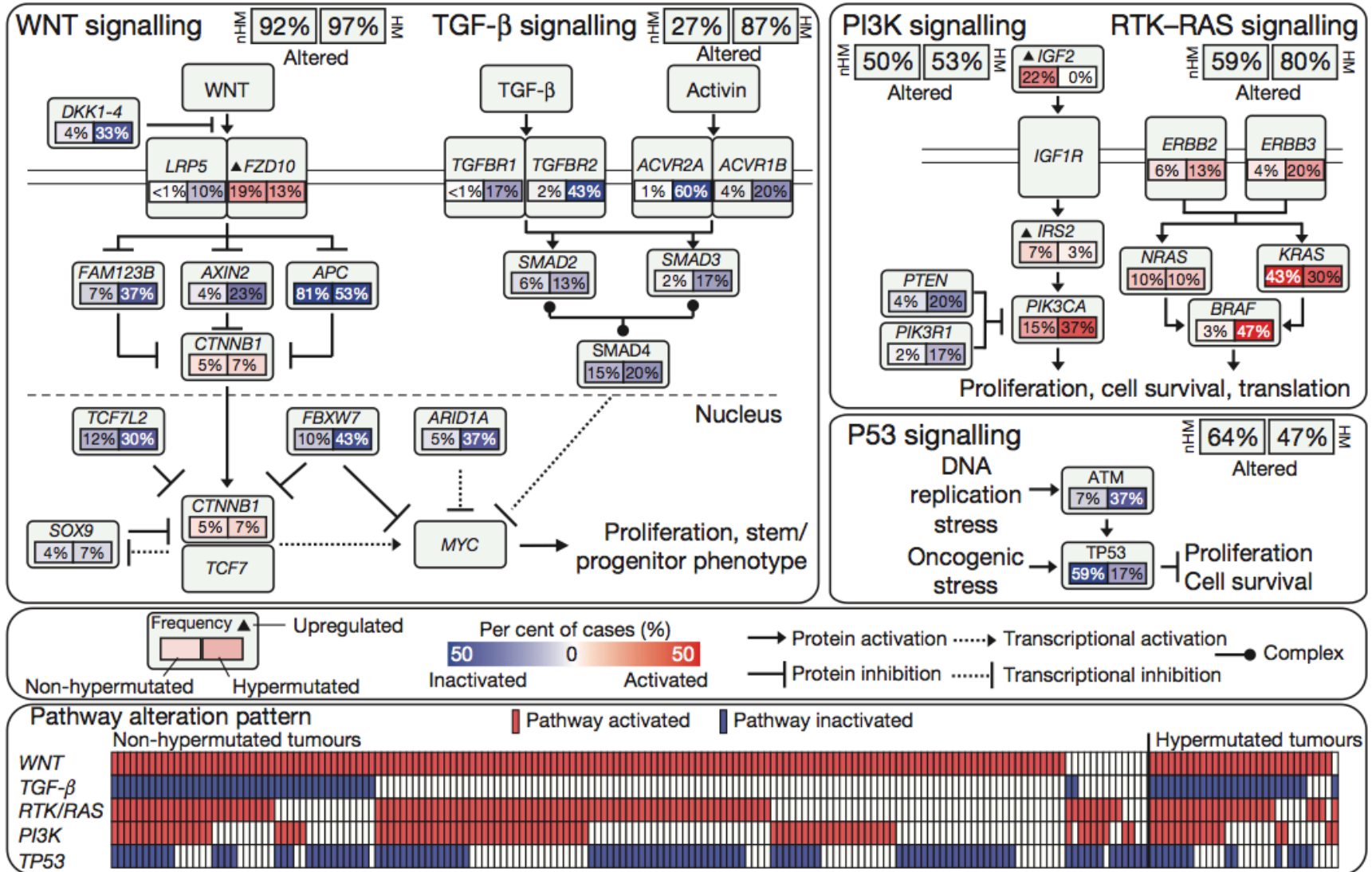


**Figure 3 | Copy-number changes and structural aberrations in CRC.**  
a, Focal amplification of 11p15.5. Segmented DNA copy-number data from single-nucleotide polymorphism (SNP) arrays and low-pass whole-genome sequencing (WGS) are shown. Each row represents a patient; amplified regions are shown in red. b, Correlation of expression levels with copy-number changes for *IGF2* and *miR-483*. c, *IGF2* amplification and overexpression are mutually

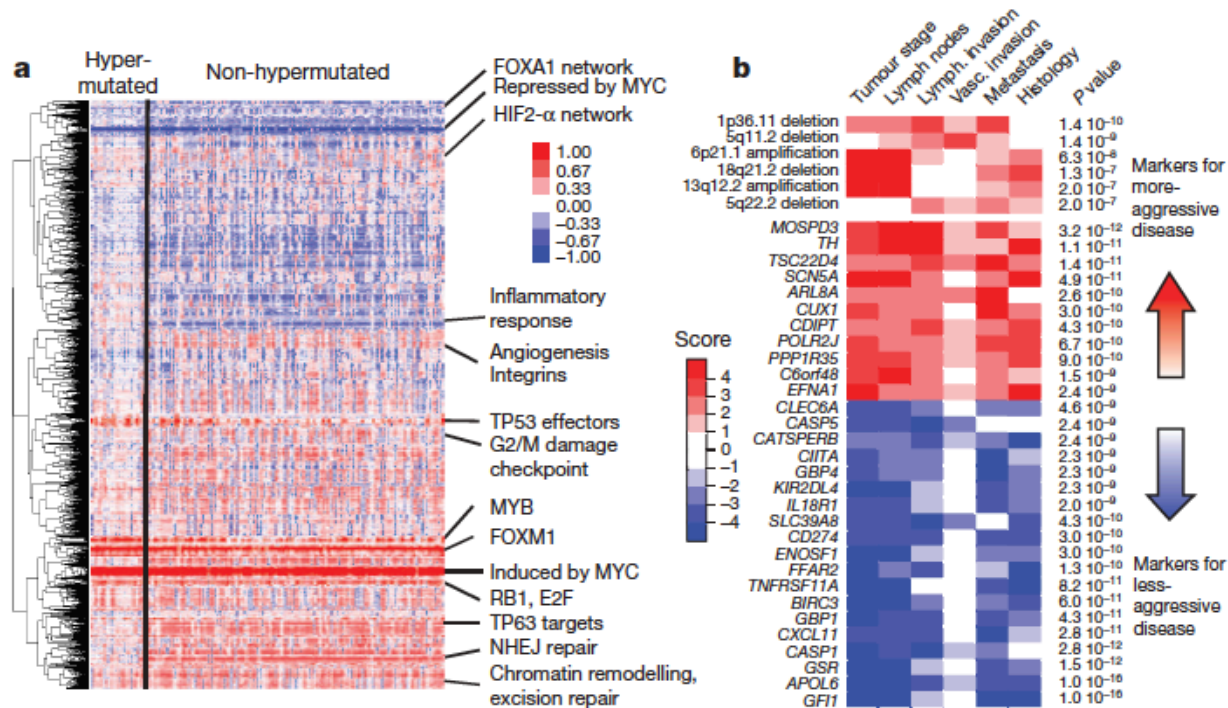
exclusive of alterations in PI3K signalling-related genes. d, Recurrent *NAV2*-*TCF7L2* fusions. The structure of the two genes, locations of the breakpoints leading to the translocation and circular representations of all rearrangements in tumours with a fusion are shown. Red line lines represent the *NAV2*-*TCF7L2* fusions and black lines represent other rearrangements. The inner ring represents copy-number changes (blue denotes loss, pink denotes gain).



# Diversity and frequency of genetic changes leading to deregulation of signaling pathways in CRC



# Integrative Pathway Analysis



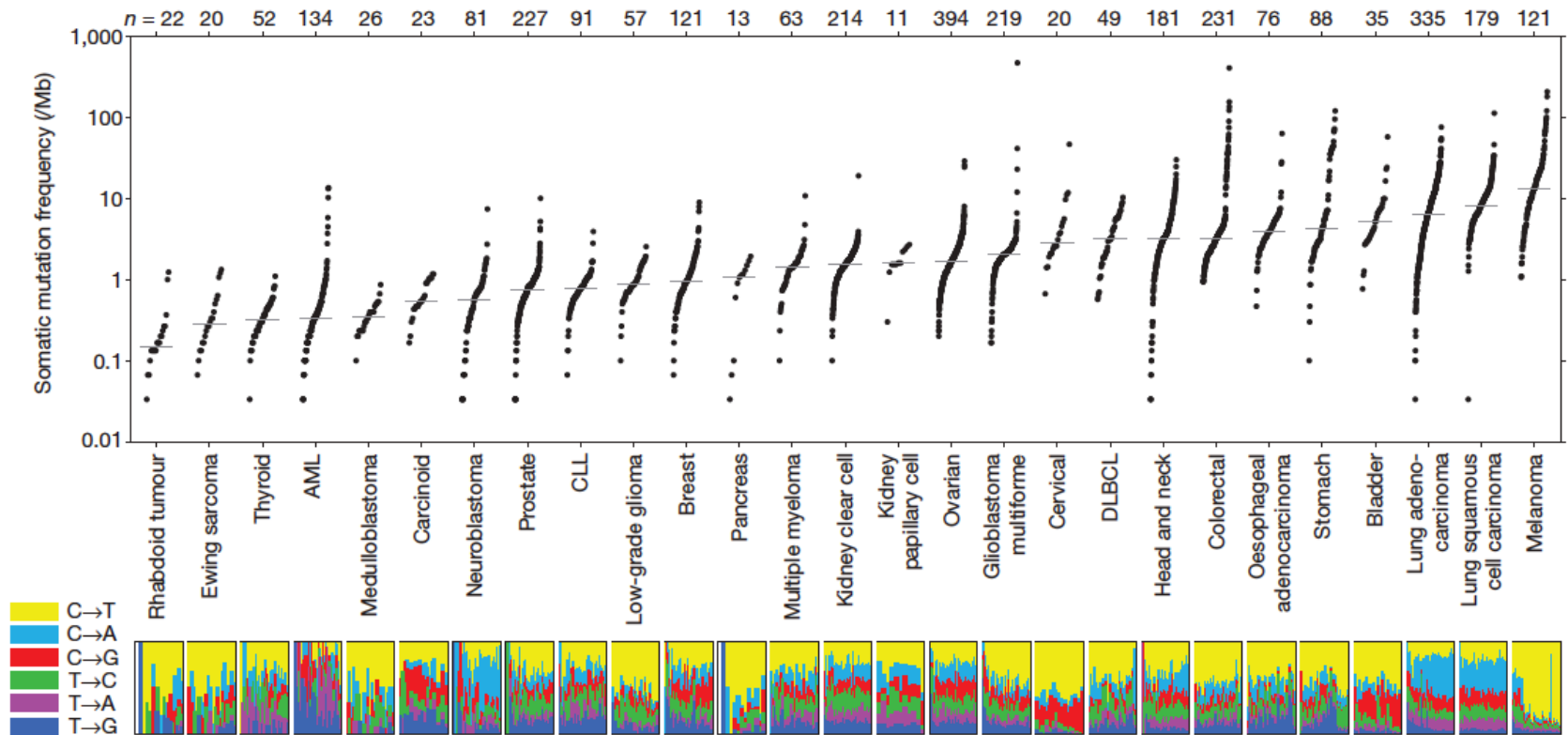
**Figure 5 | Integrative analyses of multiple data sets.** **a**, Clustering of genes and pathways affected in colon and rectum tumours deduced by PARADIGM analysis. Blue denotes under-expressed relative to normal and red denotes overexpressed relative to normal. Some of the pathways deduced by this method are shown on the right. NHEJ, non-homologous end joining. **b**, Gene-expression signatures and SCNAs associated with tumour aggression. Molecular signatures (rows) that show a statistically significant association with tumour aggressiveness according to selected clinical assays (columns) are shown in colour, with red indicating markers of tumour aggressiveness and

blue indicating the markers of less-aggressive tumours. Significance is based on the combined  $P$  value from the weighted Fisher's method, corrected for multiple testing. Colour intensity and score is in accordance with the strength of an individual clinical-molecular association, and is proportional to  $\log_{10}(P)$ , where  $P$  is the  $P$  value for that association. To limit the vertical extent of the figure, gene-expression signatures are restricted to a combined  $P$  value of  $P < 10^{-9}$  and SCNAs to  $P < 10^{-7}$ , and features are shown only if they are also significant in the subset of non-MSI-H samples (the analysis was performed separately on the full data as well as on the MSI-H and non-MSI-H subgroups).

# Pan-Cancer Analysis

## Mutational heterogeneity in cancer and the search for new cancer-associated genes

Michael S. Lawrence<sup>1\*</sup>, Petar Stojanov<sup>1,2\*</sup>, Paz Polak<sup>1,3,4\*</sup>, Gregory V. Kryukov<sup>1,3,4</sup>, Kristian Cibulskis<sup>1</sup>, Andrey Sivachenko<sup>1</sup>, Scott L. Carter<sup>1</sup>, Chip Stewart<sup>1</sup>, Craig H. Mermel<sup>1,5</sup>, Steven A. Roberts<sup>6</sup>, Adam Kiezun<sup>1</sup>, Peter S. Hammerman<sup>1,2</sup>, Aaron McKenna<sup>1,7</sup>, Yotam Drier<sup>1,3,5,8</sup>, Lihua Zou<sup>1</sup>, Alex H. Ramos<sup>1</sup>, Trevor J. Pugh<sup>1,2,3</sup>, Nicolas Stransky<sup>1,9</sup>, Elena Helman<sup>1,10</sup>, Jaegil Kim<sup>1</sup>, Carrie Sougnez<sup>1</sup>, Lauren Ambrogio<sup>1</sup>, Elizabeth Nickerson<sup>1</sup>, Erica Shefler<sup>1</sup>, Maria L. Cortés<sup>1</sup>, Daniel Auclair<sup>1</sup>, Gordon Saksena<sup>1</sup>, Douglas Voet<sup>1</sup>, Michael Noble<sup>1</sup>, Daniel DiCara<sup>1</sup>, Pei Lin<sup>1</sup>, Lee Lichtenstein<sup>1</sup>, David I. Heiman<sup>1</sup>, Timothy Fennell<sup>1</sup>, Marcin Imielinski<sup>1,5</sup>, Bryan Hernandez<sup>1</sup>, Eran Hodis<sup>1,2</sup>, Sylvan Baca<sup>1,2</sup>, Austin M. Dulak<sup>1,2</sup>, Jens Lohr<sup>1,2</sup>, Dan-Avi Landau<sup>1,2,11</sup>, Catherine J. Wu<sup>2,3</sup>, Jorge Melendez-Zajgla<sup>12</sup>, Alfredo Hidalgo-Miranda<sup>12</sup>, Amnon Koren<sup>1,3</sup>, Steven A. McCarroll<sup>1,3</sup>, Jaime Mora<sup>13</sup>, Ryan S. Lee<sup>2,3,14</sup>, Brian Crompton<sup>2,14</sup>, Robert Onofrio<sup>1</sup>, Melissa Parkin<sup>1</sup>, Wendy Winckler<sup>1</sup>, Kristin Ardlie<sup>1</sup>, Stacey B. Gabriel<sup>1</sup>, Charles W. M. Roberts<sup>2,3,14</sup>, Jaclyn A. Biegel<sup>15</sup>, Kimberly Stegmaier<sup>1,2,14</sup>, Adam J. Bass<sup>1,2,3</sup>, Levi A. Garraway<sup>1,2,3</sup>, Matthew Meyerson<sup>1,2,3</sup>, Todd R. Golub<sup>1,2,3,8</sup>, Dmitry A. Gordenin<sup>6</sup>, Shamil Sunyaev<sup>1,3,4</sup>, Eric S. Lander<sup>1,3,10</sup> & Gad Getz<sup>1,5</sup>



**Figure 1** | Somatic mutation frequencies observed in exomes from 3,083 tumour-normal pairs. Each dot corresponds to a tumour-normal pair, with vertical position indicating the total frequency of somatic mutations in the exome. Tumour types are ordered by their median somatic mutation frequency, with the lowest frequencies (left) found in haematological and paediatric tumours, and the highest (right) in tumours induced by carcinogens

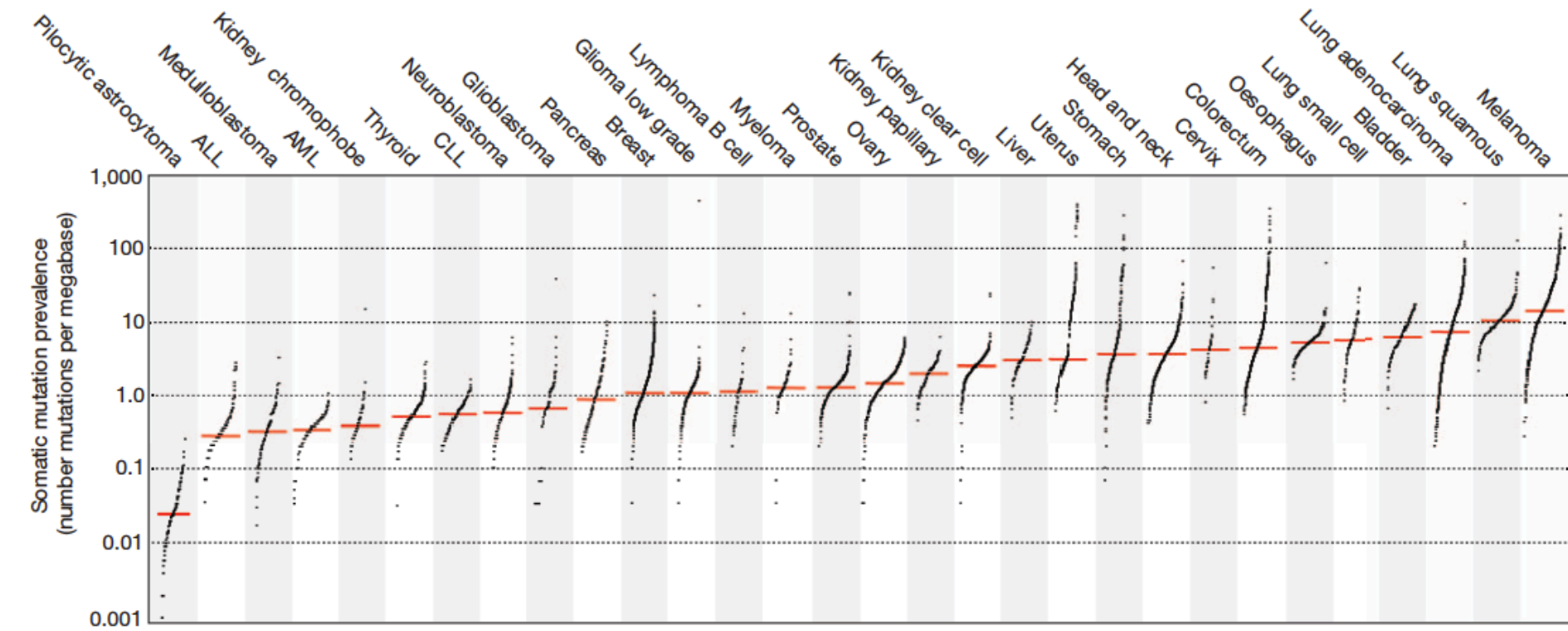
such as tobacco smoke and ultraviolet light. Mutation frequencies vary more than 1,000-fold between lowest and highest across different cancers and also within several tumour types. The bottom panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend on the left. See also Supplementary Table 2.

# Pan-Cancer Analysis

## Signatures of mutational processes in human cancer

A list of authors and their affiliations appears at the end of the paper

All cancers are caused by somatic mutations; however, understanding of the biological processes generating these mutations is limited. The catalogue of somatic mutations from a cancer genome bears the signatures of the mutational processes that have been operative. Here we analysed 4,938,362 mutations from 7,042 cancers and extracted more than 20 distinct mutational signatures. Some are present in many cancer types, notably a signature attributed to the APOBEC family of cytidine deaminases, whereas others are confined to a single cancer class. Certain signatures are associated with age of the patient at cancer diagnosis, known mutagenic exposures or defects in DNA maintenance, but many are of cryptic origin. In addition to these genome-wide mutational signatures, hypermutation localized to small genomic regions, 'kataegis', is found in many cancer types. The results reveal the diversity of mutational processes underlying the development of cancer, with potential implications for understanding of cancer aetiology, prevention and therapy.



**Figure 1 | The prevalence of somatic mutations across human cancer types.** Every dot represents a sample whereas the red horizontal lines are the median numbers of mutations in the respective cancer types. The vertical axis (log scaled) shows the number of mutations per megabase whereas the different

cancer types are ordered on the horizontal axis based on their median numbers of somatic mutations. We thank G. Getz and colleagues for the design of this figure<sup>26</sup>. ALL, acute lymphoblastic leukaemia; AML, acute myeloid leukaemia; CLL, chronic lymphocytic leukaemia.

# Pan-Cancer Analysis of TCGA Data

## Focus

Focus on TCGA Pan-Cancer Analysis

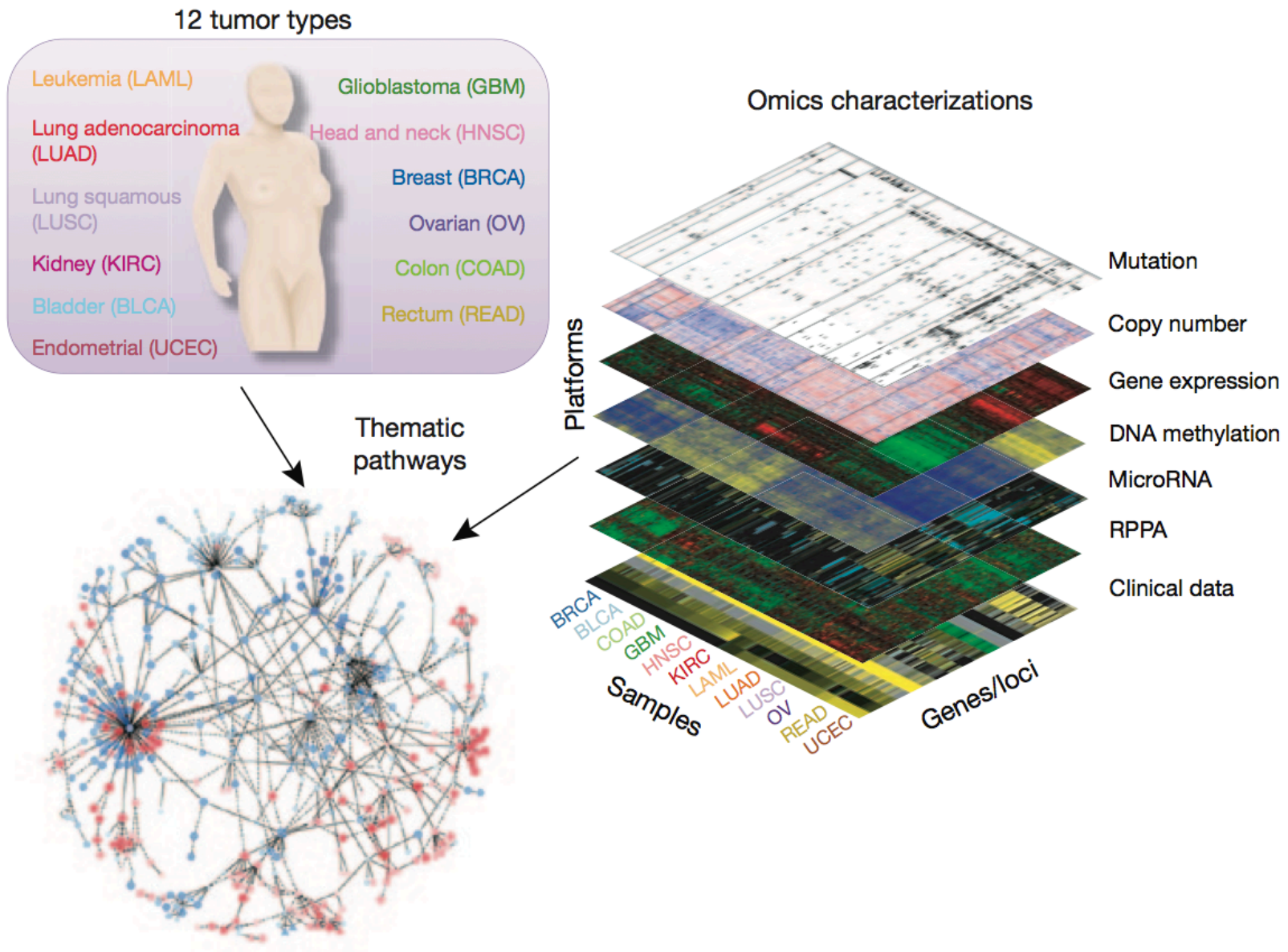


Focus issue: [October 2013](#) Volume **45**, No 10

- > [Focus contents](#)
- > [Editorial](#)
- > [Foreword](#)
- > [Analysis](#)
- > [Commentaries](#)
- > [TCGA](#)
- > [Podcast](#)

Genomic alterations in diverse cell types at different sites in the body give rise to hundreds of different forms of cancer and the ways in which these changes give rise to tumors with different biology, pathology and treatment strategies are beginning to be characterized. The Cancer Genome Atlas Research Network has catalogued the aberrations in the DNA, chromatin and RNA of the genomes of thousands of tumors relative to matched normal cellular genomes and have analyzed their epigenetic and protein consequences. Here, the Pan-cancer initiative examines the similarities and differences among the genomic and cellular alterations found in the first dozen tumor types to be profiled by TCGA. This first look across cancer offers new tools in genomics and bioinformatics and the prospect of repurposing targeted therapies directed by the molecular pathology of the tumors in addition to their clinical classification.

# Pan-Cancer Analysis of TCGA Data

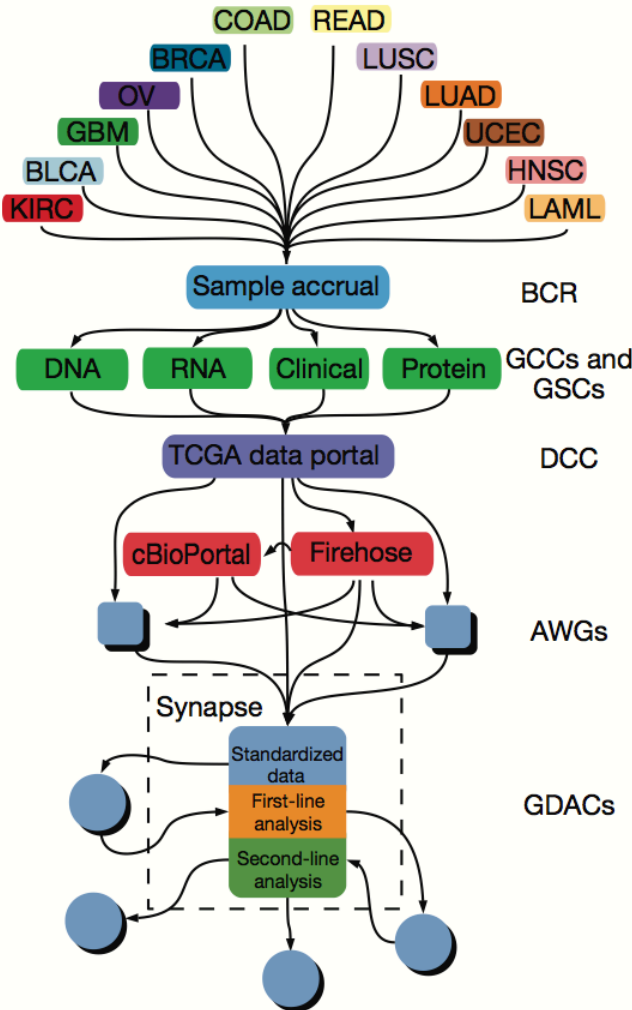


# Pan-Cancer Analysis of TCGA Data

## The Cancer Genome Atlas Pan-Cancer analysis project

The Cancer Genome Atlas Research Network<sup>1</sup>, John N Weinstein<sup>2,3</sup>, Eric A Collisson<sup>4</sup>, Gordon B Mills<sup>3</sup>, Kenna R Mills Shaw<sup>5,6</sup>, Brad A Ozenberger<sup>7</sup>, Kyle Ellrott<sup>8,9</sup>, Ilya Shmulevich<sup>10</sup>, Chris Sander<sup>11</sup> & Joshua M Stuart<sup>8,9</sup>

The Cancer Genome Atlas (TCGA) Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. The resulting rich data provide a major opportunity to develop an integrated picture of commonalities, differences and emergent themes across tumor lineages. The Pan-Cancer initiative compares the first 12 tumor types profiled by TCGA. Analysis of the molecular aberrations and their functional roles across tumor types will teach us how to extend therapies effective in one cancer type to others with a similar genomic profile.



**Table 1** Data freeze used by the Pan-Cancer project as defined on 21 December 2012

Cancer	RPPA <sup>a</sup>	DNA methylation <sup>b</sup>	Copy number <sup>c</sup>	Mutation <sup>d</sup>	microRNA <sup>e</sup>	Expression <sup>f</sup>
LUSC	195	358	345	178	332	227
READ	130	162	164	69	143	71
GBM	214	405	578	290	501	495
LAML	NA	194	198	197	187	179
HNSC	212	310	310	277	309	303
BLCA	54	126	126	99	121	96
KIRC	423	457	457	417	442	431
UCEC	200	512	511	248	497	333
LUAD	237	431	357	229	365	355
OV	332	592	577	316	454	581
BRCA	408	888	887	772	870	817
COAD	269	420	422	155	407	192
Total	2,674	4,855	4,932	3,247	4,628	4,080

Tabulated are the numbers of unique tumor samples available for each tumor type (rows) and each measurement platform (columns). NA, not available. <sup>a</sup>Reverse-phase protein arrays measuring protein and phosphoprotein abundance. <sup>b</sup>DNA methylation at CpG islands. <sup>c</sup>Microarray-based measurement of copy number. <sup>d</sup>Samples subjected to whole-exome sequencing to determine single-nucleotide and structural variants. <sup>e</sup>Sequencing of microRNAs. <sup>f</sup>RNA sequencing and microarray gene expression analysis.

# Pan-Cancer Atlas

---

<https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html>

## Welcome to the Pan-Cancer Atlas

From The Cancer Genome Atlas (TCGA) consortium, a large-scale collaboration initiated and supported by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI).

From the analysis of over 11,000 tumors from 33 of the most prevalent forms of cancer, the Pan-Cancer Atlas provides a uniquely comprehensive, in-depth, and interconnected understanding of how, where, and why tumors arise in humans. As a singular and unified point of reference, the Pan-Cancer Atlas is an essential resource for the development of new treatments in the pursuit of precision medicine.

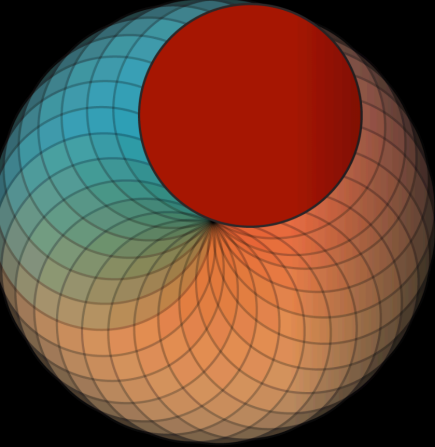
The visualization below presents the Pan-Cancer Atlas as a series of shaded rings that join together to create a beautiful, singular spectrum. Like the research itself, the full impact of this visualization is found in its cohesion. As you scroll below you will see a collection of 27 papers divided into three main categories: cell-of-origin patterns, oncogenic processes, and signaling pathways. Each category is anchored by a flagship paper providing a summary of the core findings for that topic. These are supported by companion papers that explore subtopics in depth.



# Pan-Cancer Atlas

<https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html>

## Cell-of-Origin Patterns



**Flagship Paper**

*Cell*  
**Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer**  
Peter Lind and colleagues

Comprehensive, integrated molecular analysis identifies molecular relationships across a large diverse set of human cancers, suggesting future directions for exploring clinical actionability in cancer treatment.

**Companion Papers**

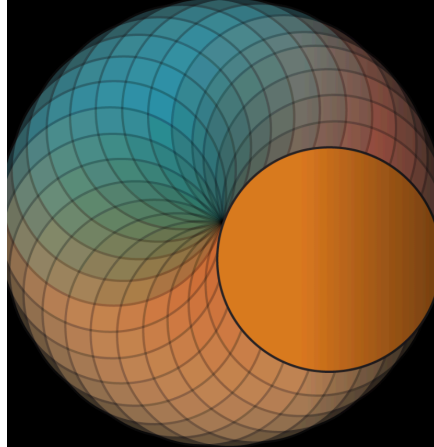
*Cell*  
**Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation**

Stemness features extracted from transcriptomic and epigenetic data from TCGA tumors reveal novel biological and clinical insight, as well as potential drug targets for anti-cancer therapies.

*Cancer Cell*  
**A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers**

By performing molecular analyses of 2,570 TCGA gynecologic (OV, UCEC, CESC, and UCS) and breast tumors, Berger et al. identify five prognostic subtypes using 16 key molecular features and propose a decision tree based on six clinically assessable features that classifies patients into the subtypes.

## Oncogenic Processes



**Flagship Paper**

*Cell*  
**Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics**  
U Ding and colleagues

A synthesized view on oncogenic processes based on PanCancer Atlas analyses highlights the complex impact of genome alterations on the signaling and multi-omic profiles of human cancers as well as their influence on tumor microenvironment.

**Companion Papers**

*Cell*  
**Pathogenic Germline Variants in 10,389 Adult Cancers**

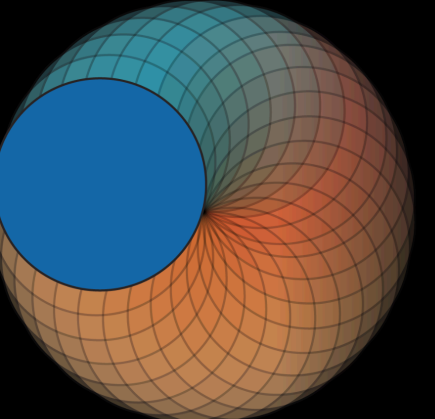
A pan-cancer analysis identifies hundreds of predisposing germline variants.

*Cell*  
**Comprehensive Characterization of Cancer Driver Genes and Mutations**

A comprehensive analysis of oncogenic driver genes and mutations in >9,000 tumors across 33 cancer types highlights the prevalence of clinically actionable cancer driver events in TCGA tumor samples.

*Cell Reports*  
**Driver Fusions and Their Implications in the Development and Treatment of Human Cancers**

## Signaling Pathways



**Flagship Paper**

*Cell*  
**Oncogenic Signaling Pathways in The Cancer Genome Atlas**  
Nikolaus Schütz and colleagues

An integrated analysis of genetic alterations in 10 signaling pathways in >9,000 tumors profiled by TCGA highlights significant representation of individual and co-occurring actionable alterations in these pathways, suggesting opportunities for targeted and combination therapies.

**Companion Papers**

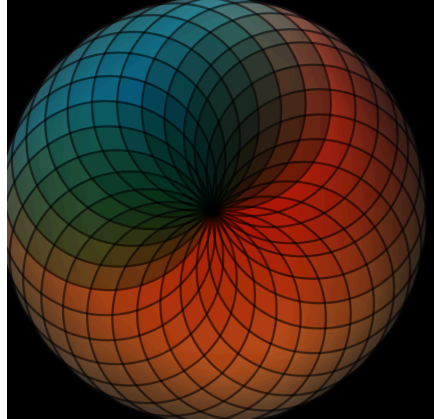
*Cell Systems*  
**Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas**

Schaub et al. present a computational study determining the frequency and extent of alterations of the MYC network across the 33 human cancers of TCGA.

*Cell Reports*  
**Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas**

Way et al. develop a machine-learning approach using Pan-Cancer Atlas data to detect Ras activation in cancer. Integrating mutation, copy number, and expression data, the authors show that their method detects Ras-activating variants in tumors and sensitivity to MEK inhibitors in cell lines.

## Resources



**National Cancer Institute**  
Access to the TCGA genomic datasets is provided through the **Genomic Data Commons**

*Cell*  
**Snapshot: TCGA-Analyzed Tumors**  
Amy Blum, Peggy Wang, and Jian C. Zorkhusen

*Cell*  
**Commentary: The Cancer Genome Atlas: Creating Lasting Value Beyond Its Data**  
Carolyn Hutter and Jean Claude Zorkhusen

*Cell*  
**Voices: The TCGA Legacy**

*Cell*  
**Editorial: Charting a Course to a Cure**  
Robert Krizger

# Tumor Heterogeneity

Cancer is not one but many diseases. It is different in each patient and continuously evolves into a progressively complex interplay of diverse tumour cells with their changing environment.

The picture that has emerged over the past few decades — especially with the advent of more sophisticated model systems and technologies — is of even greater than anticipated genetic, phenotypic and functional heterogeneity and plasticity within tumours and between primary tumours and metastases. Adding to this bewildering complexity is the heterogeneity of the tumour micro-environment, inflammatory stimuli, the immune response, mechanical stresses, therapeutic intervention and many other factors, such as diet and the microbiota. These continuously changing environmental influences affect which cancer cell subpopulations are able to survive, proliferate, spread and resist therapy.

*(Barbara Marte, Nature 2013)*

**natureINSIGHT**

TUMOUR HETEROGENEITY

19 September 2013 / Vol  
501 / Issue No 7467

## CONTENTS

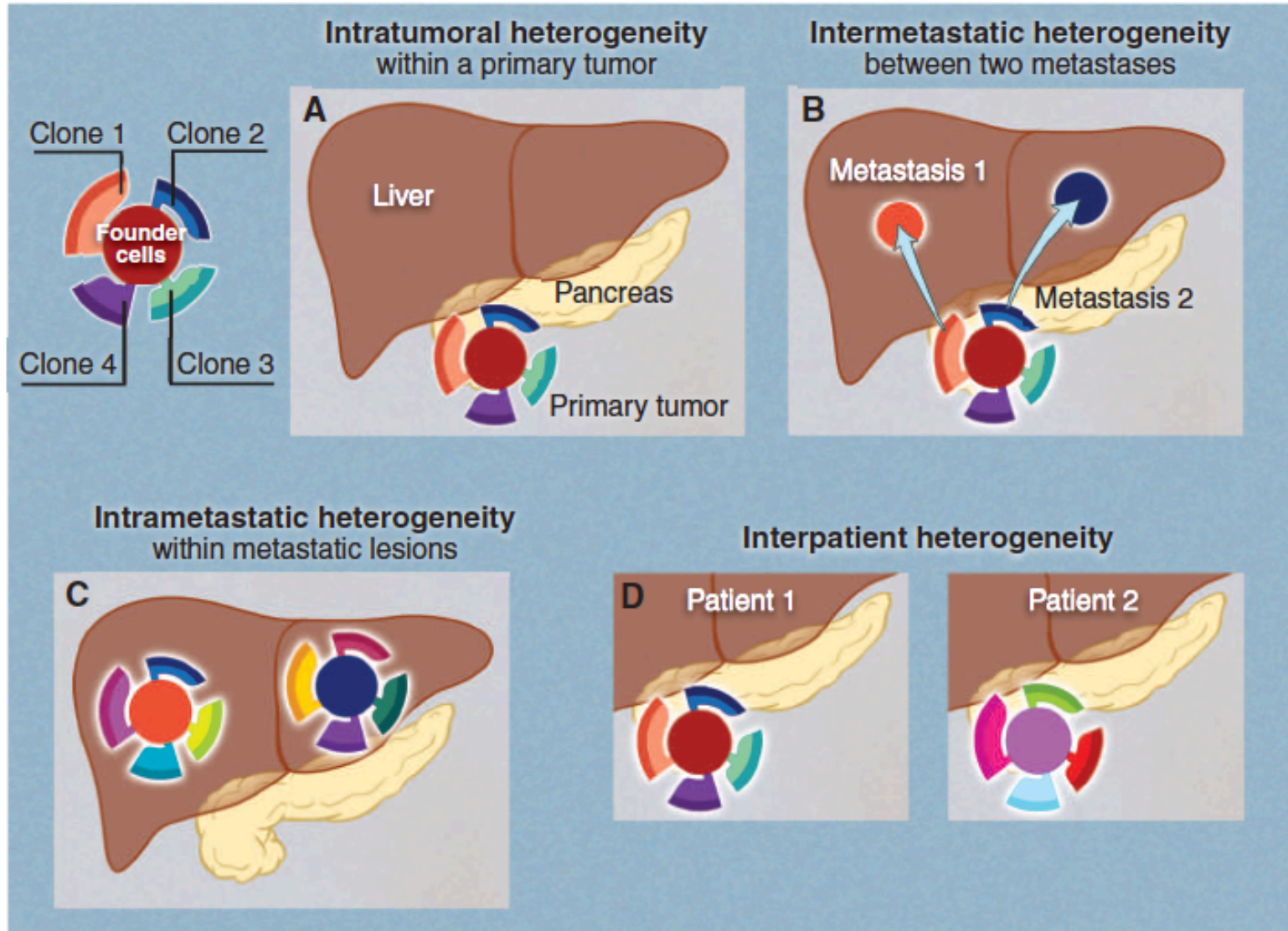
### REVIEWS

- 328** Tumour heterogeneity and cancer cell plasticity  
*Corbin E. Meacham & Sean J. Morrison*
- 338** The causes and consequences of genetic heterogeneity in cancer evolution  
*Rebecca A. Burrell, Nicholas McGranahan, Jiri Bartek & Charles Swanton*
- 346** Influence of tumour micro-environment heterogeneity on therapeutic response  
*Melissa R. Junttila & Frederic J. de Sauvage*
- 355** Tumour heterogeneity in the clinic  
*Philippe L. Bedard, Aaron R. Hansen, Mark J. Ratain & Lillian L. Siu*

### PERSPECTIVE

- 365** Selection and adaptation during metastatic cancer progression  
*Christoph A. Klein*

# Tumor Heterogeneity



# Tumor Heterogeneity

---

*The* NEW ENGLAND  
JOURNAL *of* MEDICINE

ESTABLISHED IN 1812

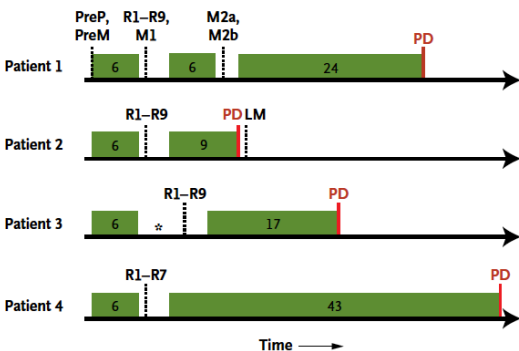
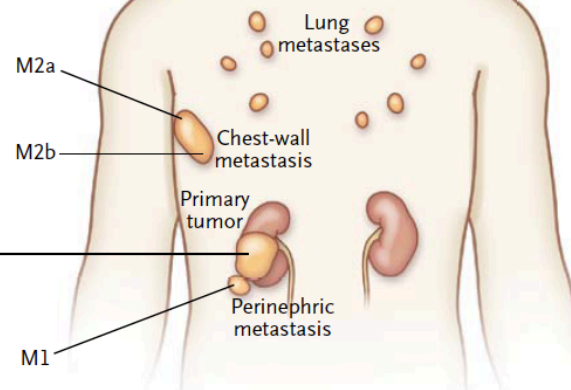
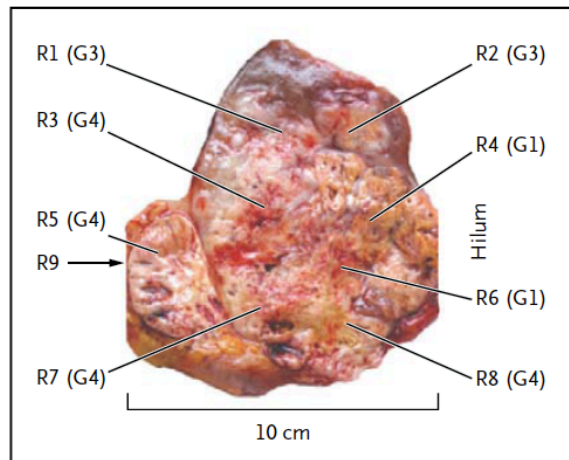
MARCH 8, 2012

VOL. 366 NO. 10

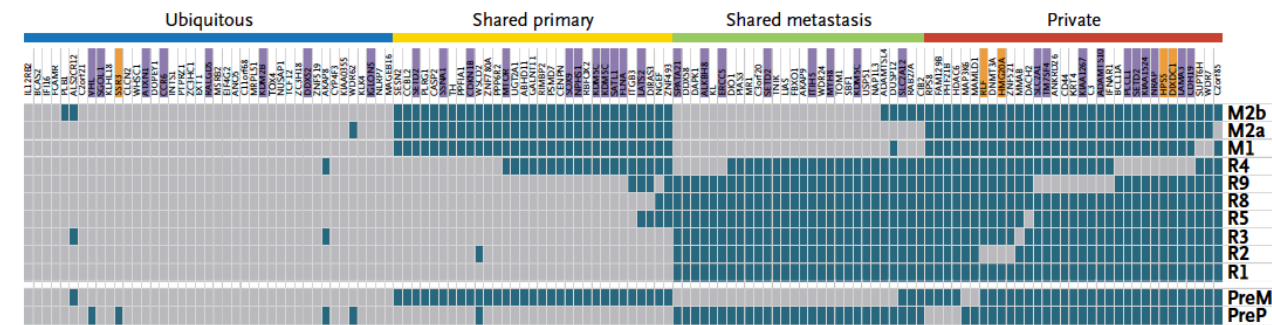
## Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing

Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horswell, M.Math., James Larkin, M.D., Ph.D., David Endesfelder, Dip.Math., Eva Gronroos, Ph.D., Pierre Martinez, Ph.D., Nicholas Matthews, B.Sc., Aengus Stewart, M.Sc., Patrick Tarpey, Ph.D., Ignacio Varela, Ph.D., Benjamin Phillipmore, B.Sc., Sharmin Begum, M.Sc., Neil Q. McDonald, Ph.D., Adam Butler, B.Sc., David Jones, M.Sc., Keiran Raine, M.Sc., Calli Latimer, B.Sc., Claudio R. Santos, Ph.D., Mahrokh Nohadani, H.N.C., Aron C. Eklund, Ph.D., Bradley Spencer-Dene, Ph.D., Graham Clark, B.Sc., Lisa Pickering, M.D., Ph.D., Gordon Stamp, M.D., Martin Gore, M.D., Ph.D., Zoltan Szallasi, M.D., Julian Downward, Ph.D., P. Andrew Futreal, Ph.D., and Charles Swanton, M.D., Ph.D.

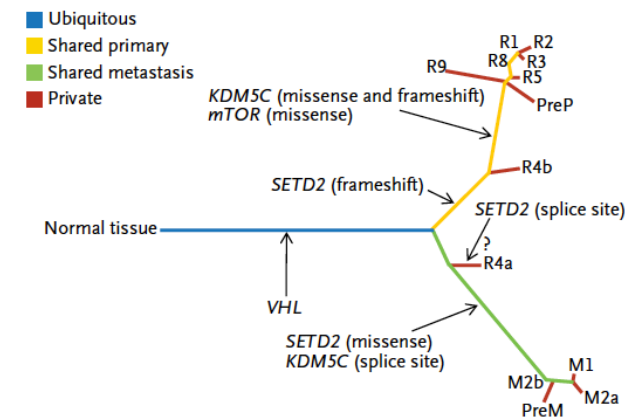
## A Biopsy Sites



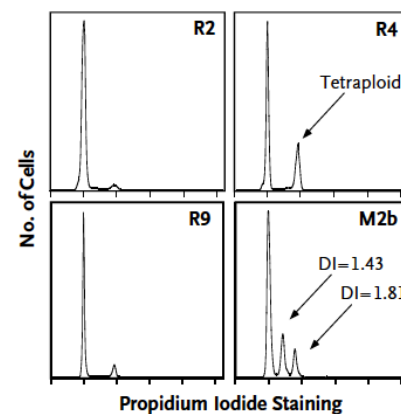
## B Regional Distribution of Mutations



## C Phylogenetic Relationships of Tumor Regions

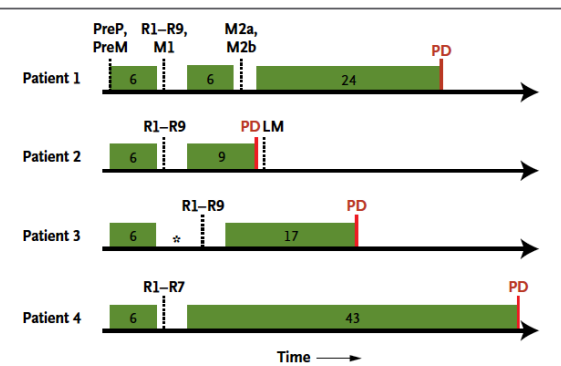


## D Ploidy Profiling



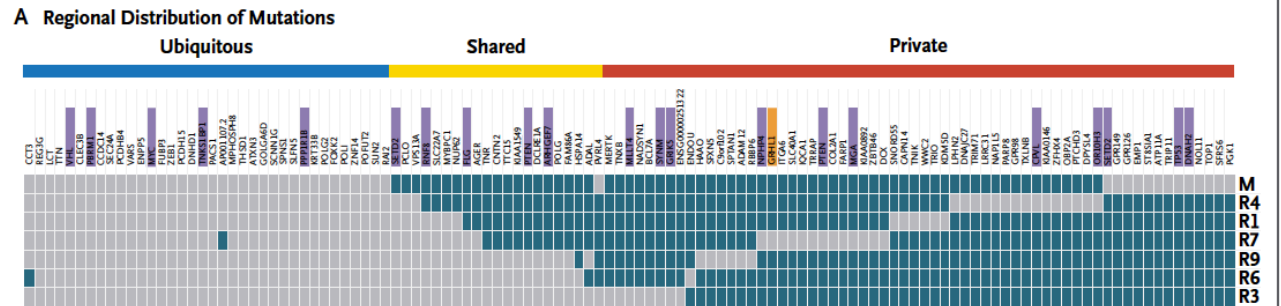
**Figure 1. Biopsy and Treatment Timelines for the Four Patients.**

Exon-capture sequencing was performed on tumor DNA from pretreatment biopsy samples of the primary tumor (PreP) and chest-wall metastasis (PreM), primary-tumor regions of the nephrectomy specimen (R1 to R9), a perinephric metastasis in the nephrectomy specimen (M1), and two regions of the excised chest-wall metastasis (M2a and M2b). LM denotes liver metastasis, and PD progressive disease. Green boxes indicate periods of everolimus treatment, with the treatment duration provided in weeks. Dotted lines indicate time points of biopsies, and the asterisk indicates a delay in nephrectomy because of toxicity.

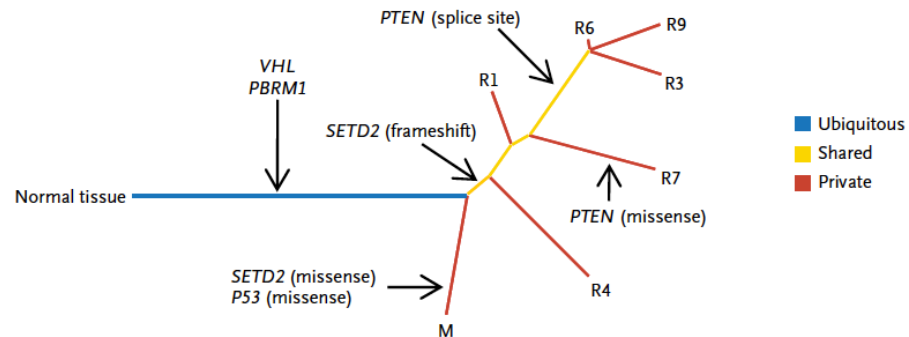


**Figure 1. Biopsy and Treatment Timelines for the Four Patients.**

Exon-capture sequencing was performed on tumor DNA from pretreatment biopsy samples of the primary tumor (PreP) and chest-wall metastasis (PreM), primary-tumor regions of the nephrectomy specimen (R1 to R9), a perinephric metastasis in the nephrectomy specimen (M1), and two regions of the excised chest-wall metastasis (M2a and M2b). LM denotes liver metastasis, and PD progressive disease. Green boxes indicate periods of everolimus treatment, with the treatment duration provided in weeks. Dotted lines indicate time points of biopsies, and the asterisk indicates a delay in nephrectomy because of toxicity.



**B Phylogenetic Relationships of Tumor Regions**



**Figure 4. Genetic Intratumor Heterogeneity and Phylogeny in Patient 2.**

Panel A shows the regional distribution of somatic mutations detected by exome sequencing in a heat map, with gray indicating the presence of a mutation and dark blue the absence of a mutation. The color bars above the heat map indicate classification of mutations according to whether they are ubiquitous, shared by primary-tumor regions, or unique to the region (private). For gene names, purple indicates that the mutation was validated, and orange indicates that the validation of the mutation failed. Panel B shows phylogenetic relationships of the tumor regions. Branch lengths are proportional to the number of somatic mutations separating the branching points. Potential driver mutations were acquired by the indicated genes in the branch (arrows).

# Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions

ORIGINAL ARTICLE



## Innate Genetic Evolution of Lung Cancers and Spatial Heterogeneity: Analysis of Treatment-Naïve Lesions



Kenichi Suda, MD, PhD,<sup>a,b</sup> Jihye Kim, PhD,<sup>a</sup> Isao Murakami, MD, PhD,<sup>c</sup> Leslie Rozeboom, MS,<sup>a</sup> Masaki Shimoji, MD, PhD,<sup>b</sup> Shigeki Shimizu, MD, PhD,<sup>d</sup> Christopher J. Rivard, PhD,<sup>a</sup> Tetsuya Mitsudomi, MD, PhD,<sup>b</sup> Aik-Choon Tan, PhD,<sup>a,\*</sup> Fred R. Hirsch, MD, PhD<sup>a</sup>

<sup>a</sup>Division of Medical Oncology, University of Colorado Anschutz Medical Campus, Aurora, Colorado  
<sup>b</sup>Division of Thoracic Surgery, Department of Surgery, Kindai University Faculty of Medicine, Osaka-Sayama, Japan  
<sup>c</sup>Department of Respiratory Medicine, Higashi-Hiroshima Medical Center, Higashi-Hiroshima, Japan  
<sup>d</sup>Department of Pathology, Kindai University Faculty of Medicine, Osaka-Sayama, Japan

Received 14 February 2018; revised 12 May 2018; accepted 22 May 2018  
 Available online - 19 June 2018

### ABSTRACT

**Introduction:** Data regarding the pre-treatment inter-tumor heterogeneity of potential biomarkers in advanced-stage lung cancers is limited. A finding of such heterogeneity between primary and metastatic lesions would prove valuable to determine if a metastatic lesion can be a surrogate for the primary tumor, as more biomarkers will likely be used in the future to inform treatment decisions.

**Methods:** We performed RNA sequencing to analyze inter-tumor heterogeneity in 30 specimens (primary tumors, intrathoracic, and extrathoracic metastatic lesions) obtained from five treatment-naïve lung cancer patients.

**Results:** The global unsupervised clustering analysis showed that the lesions clustered at the individual patient level rather than on the metastatic sites, suggesting that the characteristics of specific tumor cells have a greater impact on the gene expression signature than the microenvironment in which the metastasis develops. The mutational and transcriptional data highlight the presence of inter-tumor heterogeneity showing that the primary tumors are usually distinct from metastatic lesions. Through a comparison between metastatic lesions and the primary tumors, we observed that pathways related to cell proliferation were upregulated, whereas immune-related pathways were downregulated in metastatic lesions.

**Conclusion:** These data not only provide insight into the evolution of lung cancers, but also imply possibilities and limitations of biomarker-based treatment in lung cancers.

© 2018 International Association for the Study of Lung Cancer. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Keywords:** Biomarkers; Tumor heterogeneity; RNA sequencing; Autopsy; Immune-related markers; RET fusion

### Introduction

The recent development of personalized molecular targeted therapies in lung cancer based on genetic aberrations of tumor cells has dramatically improved the efficacy of systemic therapies and prolonged patient

\*Corresponding author.

Drs. Suda and Kim contributed equally to this work.

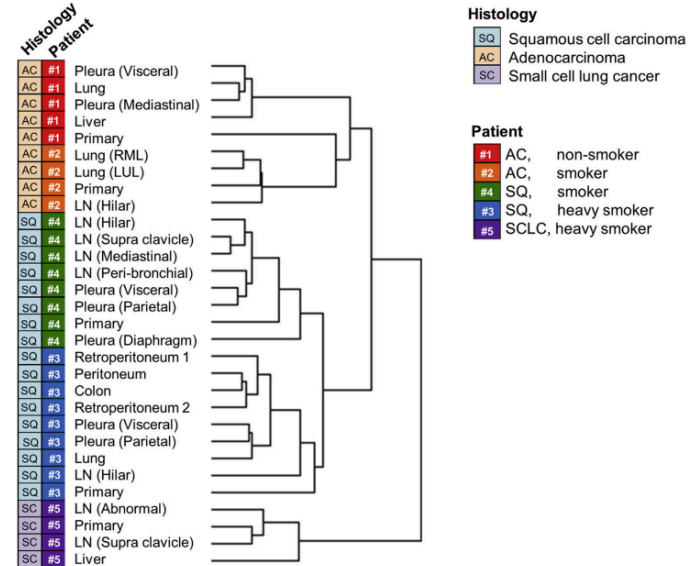
**Disclosure:** Dr. Mitsudomi has received grants from Astra Zeneca and Chugai; and has received personal fees from Astra Zeneca, Chugai, Boehringer Ingelheim, Pfizer, Roche, and Clovis Oncology. Dr. Hirsch has received grants from Genentech/Roche, Bristol-Myers Squibb, Lilly, Bayer, Amgen, and Ventana/Roche; and has received personal fees from Genentech/Roche, Pfizer, Bristol-Myers Squibb, Lilly, Merck, Ventana/Roche, Novartis, and Abbvie. The remaining authors declare no conflict of interest.

Address for correspondence: Aik-Choon Tan, PhD, Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus, MS 8117, 12801 E. 17th Ave., Aurora, Colorado 80045. E-mail: aikchoon.tan@ucdenver.edu

© 2018 International Association for the Study of Lung Cancer. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

ISSN: 1556-0864

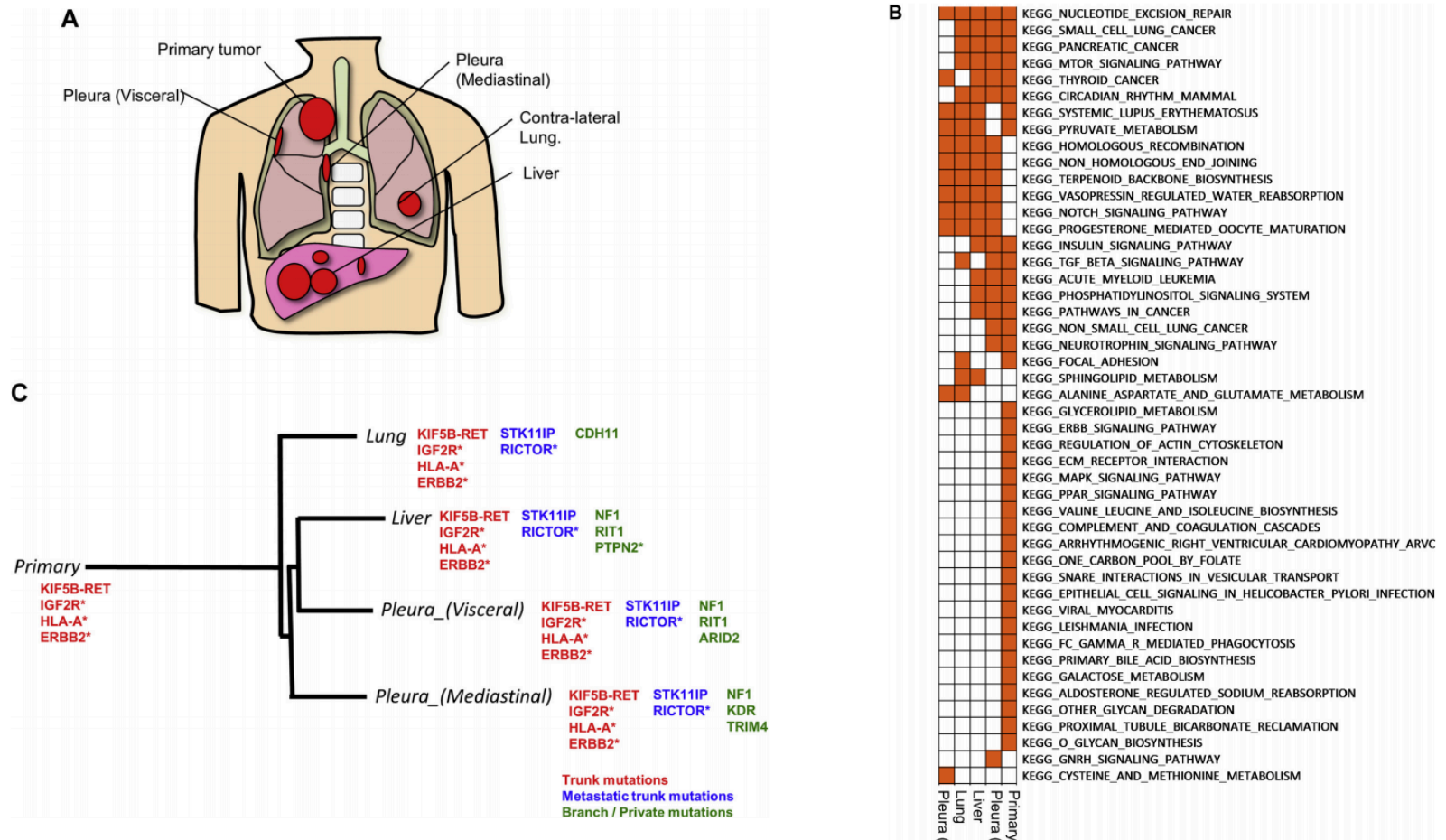
<https://doi.org/10.1016/j.jtho.2018.05.039>



**Figure 1.** The global unsupervised clustering analysis of expression data of all lesions analyzed. The left and right rows indicate the histology and the individual patient, respectively. LN, lymph node metastases; RML, right middle lobe of the lung; LUL, left upper lobe of the lung; AC, adenocarcinoma; SQ, squamous cell carcinoma.



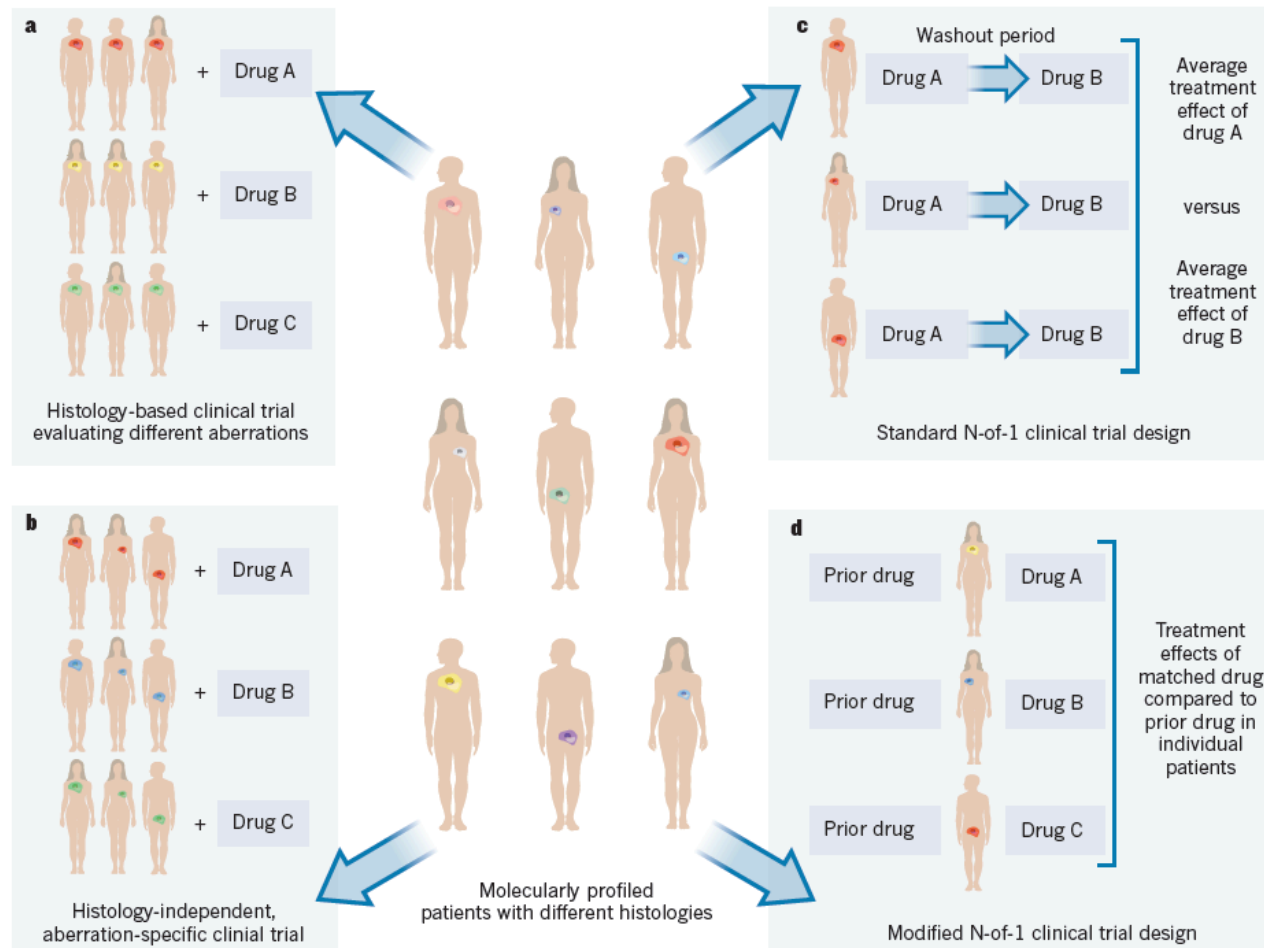
# Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions



**Figure 2.** Mutational and pathway analyses of lesions obtained from a never-smoking lung adenocarcinoma patient (case 1). (A) Geographic locations of analyzed lesions are shown. (B) Pathway analysis was performed with Gene Set Enrichment Analysis based on MSigDB C2 curated KEGG gene sets. (Dysregulated pathways in all lesions are summarized in Supplementary Table 1, and are excluded from the figure.) (C) Representative mutations identified in case 1. *KIF5B-RET* fusion was also identified in all lesions, but not in noncancerous tissue, using TopHat-Fusion on the RNA-seq. Insulin-like growth factor 2 receptor (*IGF2R*), major histocompatibility complex, class I, A (*HLA-A*), and erb-b2 receptor tyrosine kinase 2 (*ERBB2*) mutations were also identified as trunk mutations, whereas *NF1* mutation was present only in some metastatic lesions. The phylogenetic tree was constructed using a Neighbor joining method implemented in the R phangom package, based on somatic and deleterious mutations in reliably expressed genes. The genes with asterisk indicate that somatic/deleterious mutations in those genes were identified in lesions obtained from the other lung adenocarcinoma patient (case 2).



# Implications in Clinical Trial Design



**Figure 1 | Clinical-trial design frameworks.** In a population of molecularly profiled patients who have tumours of different histologies (shown by position of tumour) and molecular aberrations (shown as different colours), the framework for a clinical trial can take a number of forms. **a**, Histology-based clinical trials evaluate different molecular aberrations by enrolling patients with the same tumour histology but who harbour different aberrations, and match groups of patients to different drugs. **b**, Histology-independent, aberration-specific clinical trials, or 'basket' trials, enrol patients with different tumour histologies but who

harbour the same or related molecular aberrations, and match drugs to the aberration specific or related groups. **c**, Standard N-of-1 trials randomly assign patients to different drugs in different sequential orders, with washout periods between drugs to minimize crossover effects. At completion, the individual effect of each drug and the average effects of each drug across individuals can be analysed. **d**, Modified N-of-1 trials use each patient as his or her own control and compare the treatment effect of the current matched drug with that of the most recent earlier drug. (Bedard et al Nature 2013)

# New Clinical Trial Designs

The NEW ENGLAND JOURNAL of MEDICINE

## REVIEW ARTICLE

### THE CHANGING FACE OF CLINICAL TRIALS

Jeffrey M. Drazen, M.D., David P. Harrington, Ph.D., John J.V. McMurray, M.D., James H. Ware, Ph.D., and Janet Woodcock, M.D., Editors

## Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both

Janet Woodcock, M.D., and Lisa M. LaVange, Ph.D.

**H**IGH-QUALITY EVIDENCE IS WHAT WE USE TO GUIDE MEDICAL PRACTICE. The standard approach to generating this evidence — a series of clinical trials, each investigating one or two interventions in a single disease — has become ever more expensive and challenging to execute. As a result, important clinical questions go unanswered. The conduct of “precision medicine” trials to evaluate targeted therapies creates challenges in recruiting patients with rare genetic subtypes of a disease. There is also increasing interest in performing mechanism-based trials in which eligibility is based on criteria other than traditional disease definitions. The common denominator is a need to answer more questions more efficiently and in less time.

A methodologic innovation responsive to this need involves coordinated efforts to evaluate more than one or two treatments in more than one patient type or disease within the same overall trial structure.<sup>1-4</sup> Such efforts are referred to as master protocols, defined as one overarching protocol designed to answer multiple questions. Master protocols may involve one or more interventions in multiple diseases or a single disease, as defined by current disease classification, with multiple interventions, each targeting a particular biomarker-defined population or disease subtype. Included under this broad definition of a master protocol are three distinct entities: umbrella, basket, and platform trials (Table 1 and Figs. 1 and 2). All constitute a collection of trials or substudies that share key design components and operational aspects to achieve better coordination than can be achieved in single trials designed and conducted independently.

A master protocol may involve direct comparisons of competing therapies or be structured to evaluate, in parallel, different therapies relative to their respective controls. Some take advantage of existing infrastructure to capitalize on similarities among trials, whereas others involve setting up a new trial network specific to the master protocol. All require intensive pretrial discussion among sponsors contributing therapies for evaluation and parties involved in the conduct and governance of the trials to ensure that issues surrounding data use, publication rights, and the timing of regulatory submissions are addressed and resolved before the start of the trial.

**Table 1.** Types of Master Protocols.

Type of Trial	Objective
Umbrella	To study multiple targeted therapies in the context of a single disease
Basket	To study a single targeted therapy in the context of multiple diseases or disease subtypes
Platform	To study multiple targeted therapies in the context of a single disease in a perpetual manner, with therapies allowed to enter or leave the platform on the basis of a decision algorithm

From the Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD. Address reprint requests to Dr. LaVange at the Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, Food and Drug Administration, 10903 New Hampshire Blvd., Silver Spring, MD 20993, or at [lisa.lavange@fda.hhs.gov](mailto:lisa.lavange@fda.hhs.gov).

N Engl J Med 2017;377:62-70.

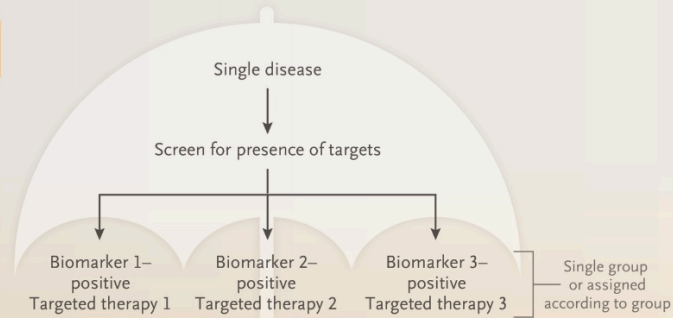
DOI: 10.1056/NEJMra1510062

Copyright © 2017 Massachusetts Medical Society.

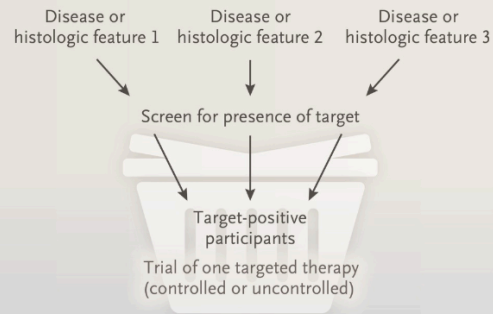
# New Clinical Trial Designs



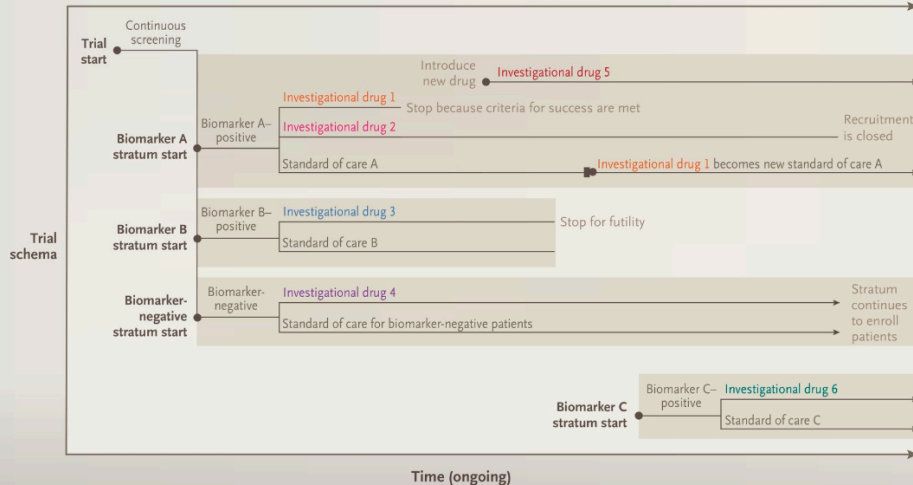
**Umbrella trial**



**Basket trial**



**Trial events**



# New Clinical Trial Designs

**Table 2. Examples of Master Protocols in Cancer.\***

Trial	Description	Design	Drug or Drugs	Disease and Target	Study Population	End Points
B2225 <sup>6</sup>	Basket trial to determine cancers responsive to imatinib	Phase 2, multicenter, open-label, noncomparative trial	Single: imatinib (400 or 800 mg per day)	40 cancers (solid tumors and hematologic cancers) with activation of imatinib target kinases	186 patients $\geq 15$ yr of age	Tumor response (SWOG criteria and investigator's assessment)
<i>BRAF</i> V600 <sup>7</sup>	Basket trial to evaluate the efficacy of vemurafenib in nonmelanoma cancers	Early phase 2, multicenter, open-label, noncomparative, adaptive trial using Simon's two-stage design	Vemurafenib monotherapy or (in some patients with colorectal cancer) vemurafenib plus cetuximab	Multiple nonmelanoma cancers with <i>BRAF</i> V600 mutations; eight tumor-specific cohorts plus an "all others" cohort	122 adults ( $\geq 18$ yr of age)	Response rate (assessed by investigators according to RECIST or IMWG criteria) at wk 8
NCI-Match <sup>8</sup>	Umbrella trial to determine whether treating cancers according to molecular abnormalities is effective	Exploratory, multicenter, noncomparative trial	Multiple: 30 treatments (as of May 2016), both FDA-approved and investigational, that target gene abnormalities	Advanced solid tumor, lymphoma, or myeloma; DNA sequencing for actionable mutations	35 adults planned per substudy; pediatric study to begin in 2017	Tumor response (primary) and progression-free survival
BATTLE-1 <sup>9</sup>	Umbrella trial to evaluate targeted therapies in chemotherapy-refractory NSCLC	Phase 2, single-center, comparative, adaptive randomization trial	Multiple: three monotherapies (erlotinib, vandetanib, and sorafenib) and one combination (erlotinib plus bevacizumab)	Advanced NSCLC; targets included <i>EGFR</i> mutation, <i>KRAS/BRAF</i> mutation, VEGF expression, and RXRs/CyclinD1 expression	255 adults in whom $\geq 1$ chemotherapy regimen had failed	Complete or partial response or stable disease according to RECIST criteria at wk 8 (primary), progression-free survival, overall survival, and toxicity
I-SPY 2 <sup>10-12</sup>	Adaptive platform trial to identify treatment regimens for locally advanced breast cancer in the context of neoadjuvant therapy on the basis of biomarker signatures	Phase 2, multicenter, comparative, adaptive randomization trial	Multiple: standard chemotherapy and five new drugs (initially) as add-on to chemotherapy; 12 treatments tested to date, with latest (pamiparib) added October 2016	Early, high-risk breast cancer; three biomarkers (hormone-receptor status, HER2 status, and MammaPrint risk score) define eight genetic subgroups	1920 women (estimated) with invasive tumor $\geq 2.5$ cm in diameter	Pathological complete response
Lung-MAP <sup>13-15</sup>	Master protocol to evaluate biomarker-matched therapies in rare squamous-cell subsets of NSCLC	Phase 2-3 comparative trial	Multiple: four investigational drugs plus one therapy for no-match control group (initially); three investigational drugs remain	Squamous-cell NSCLC; multiple targets (four molecular targets initially; three remain)	100-170 patients planned for phase 2 (40 are now enrolled); 300-400 planned for phase 3	Objective response rate, progression-free survival, and overall survival

\* BATTLE-1 denotes Biomarker-Integrated Approaches of Targeted Therapy for Lung Cancer Elimination 1, IMWG International Myeloma Working Group, I-SPY 2 Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis 2, Lung-MAP Lung Master Protocol, NCI-MATCH National Cancer Institute Molecular Analysis for Therapy Choice, NSCLC non-small-cell lung cancer, RECIST Response Evaluation Criteria in Solid Tumors, and SWOG Southwest Oncology Group.

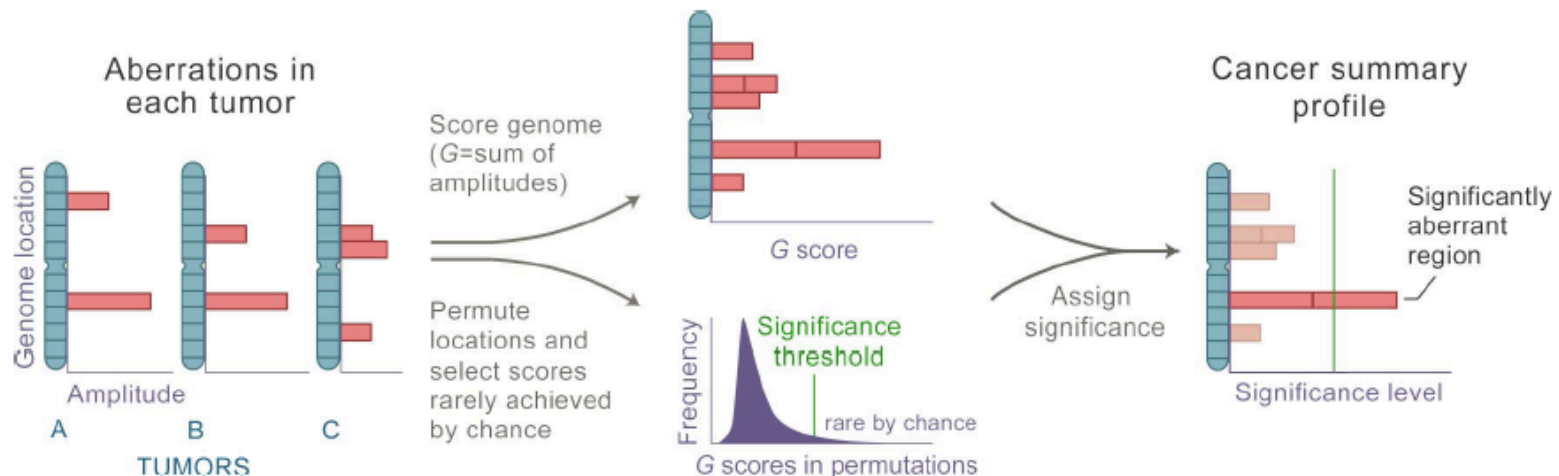
# Bioinformatics Tools for Mining and Visualizing Cancer Genomics Data

---

- “Driver” Mutations
  - GISTIC – Copy number variation
  - MutSig – DNA somatic mutation
  - PARADIGM – pathway inference for individual patients
- Analysis Pipeline
  - Firehose
  - Nozzle
- Cancer Genome Browser
  - UCSC Cancer Genome Browser
- Data Portal
  - TCGA Data Portal
  - ICGC Data Portal
  - cBio Cancer Genome Portal

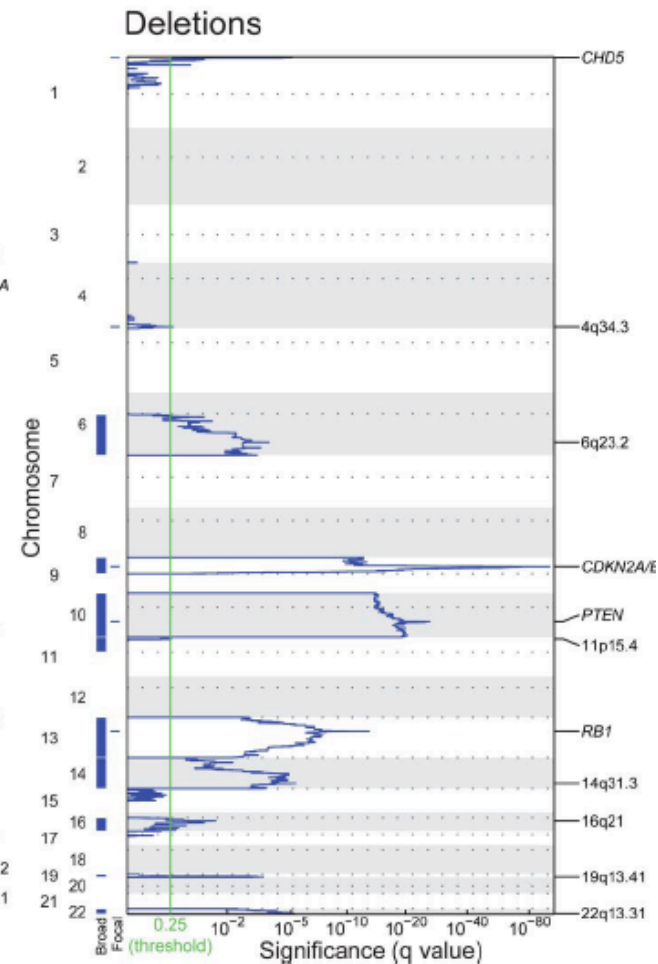
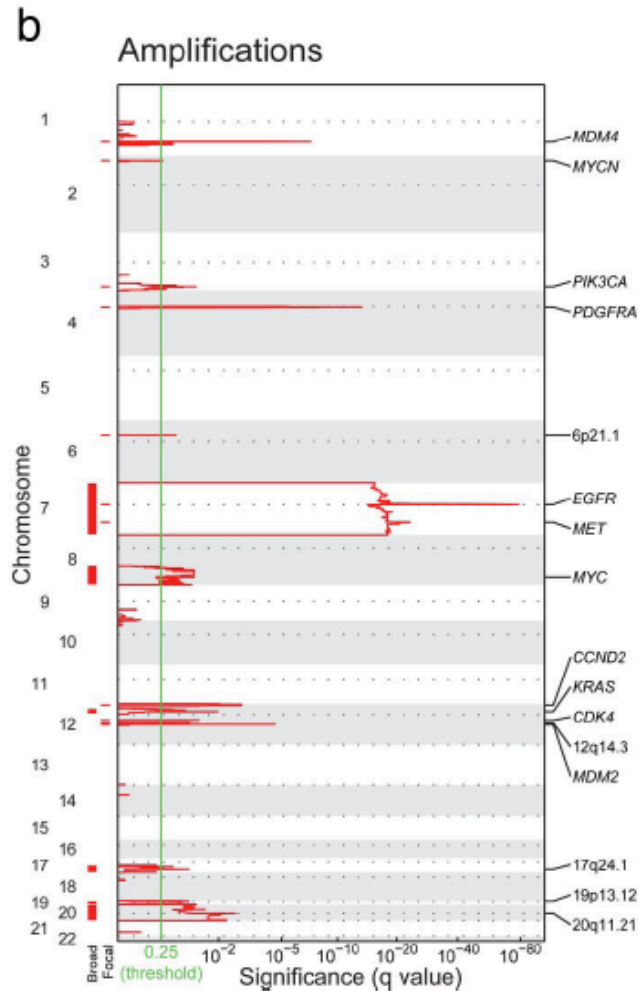
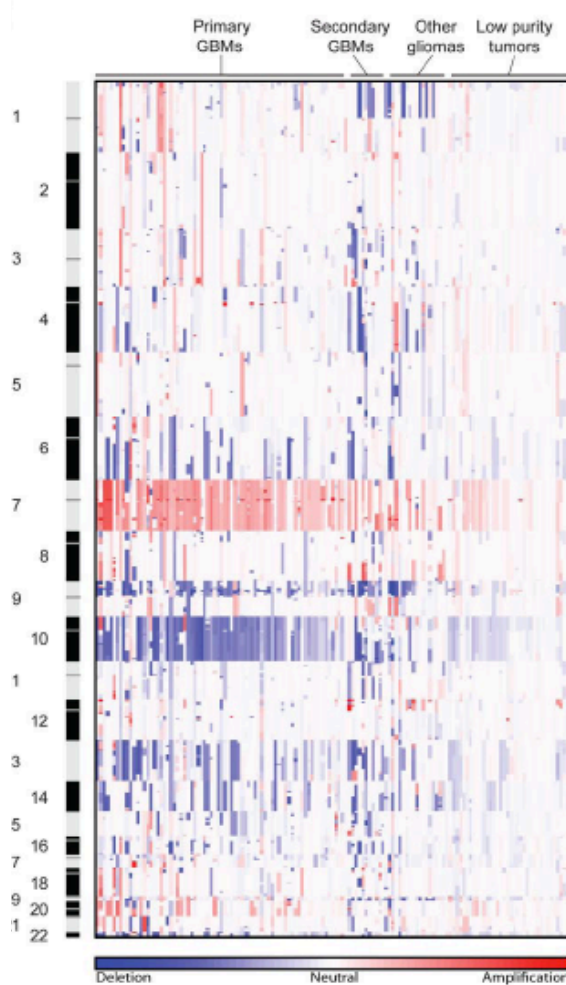
# Genomic Identification of Significant Targets in Cancer (GISTIC)

- Statistical method for identifying regions of aberration that are more likely to drive cancer pathogenesis. The method identifies those regions of the genome that are aberrant more often than would be expected by chance, with greater weight given to high amplitude events (high-level copy-number gains or homozygous deletions) that are less likely to represent random aberrations.



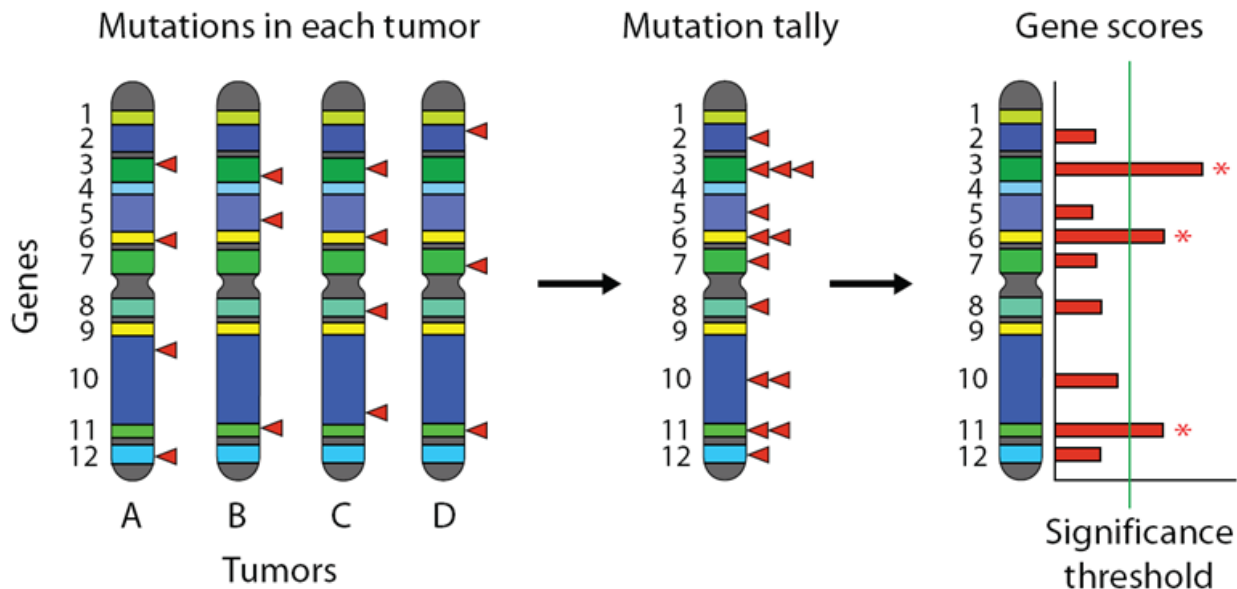
**Fig. 1.** Overview of the GISTIC method. After identifying the locations and, in the case of copy-number alterations, magnitudes (as  $\log_2$  signal intensity ratios) of chromosomal aberrations in multiple tumors (*Left*), GISTIC scores each genomic marker with a  $G$  score that is proportional to the total magnitude of aberrations at each location (*Upper Center*). In addition, by permuting the locations in each tumor, GISTIC determines the frequency with which a given score would be attained if the events were due to chance and therefore randomly distributed (*Lower Center*). A significance threshold (green line) is determined such that significant scores are unlikely to occur by chance alone. Alterations are deemed significant if they occur in regions that surpass this threshold (*Right*). For more details see [SI Text](#).

# Genomic Identification of Significant Targets in Cancer (GISTIC)



# MutSig (and variants) to identify “Mutation Significance” from Sequencing Data

- The input data is lists of mutations (and indels) from a set of samples (patients) that were subjected to DNA sequencing, as well as information about how much territory was covered in the sequencing.
- MutSig builds a model of the background mutation processes that were at work during formation of the tumors, and it analyzes the mutations of each gene to identify genes that were mutated more often than expected by chance, given the background model.





## Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM

Charles J. Vaske<sup>1,†</sup>, Stephen C. Benz<sup>2,†</sup>, J. Zachary Sanborn<sup>2</sup>, Dent Earl<sup>2</sup>, Christopher Szeto<sup>2</sup>, Jingchun Zhu<sup>2</sup>, David Haussler<sup>1,2</sup> and Joshua M. Stuart<sup>2,\*</sup>

<sup>1</sup>Howard Hughes Medical Institute and <sup>2</sup>Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, UC Santa Cruz, CA, USA

### ABSTRACT

**Motivation:** High-throughput data is providing a comprehensive view of the molecular changes in cancer tissues. New technologies allow for the simultaneous genome-wide assay of the state of genome copy number variation, gene expression, DNA methylation and epigenetics of tumor samples and cancer cell lines. Analyses of current data sets find that genetic alterations between patients can differ but often involve common pathways. It is therefore critical to identify relevant pathways involved in cancer progression and detect how they are altered in different patients.

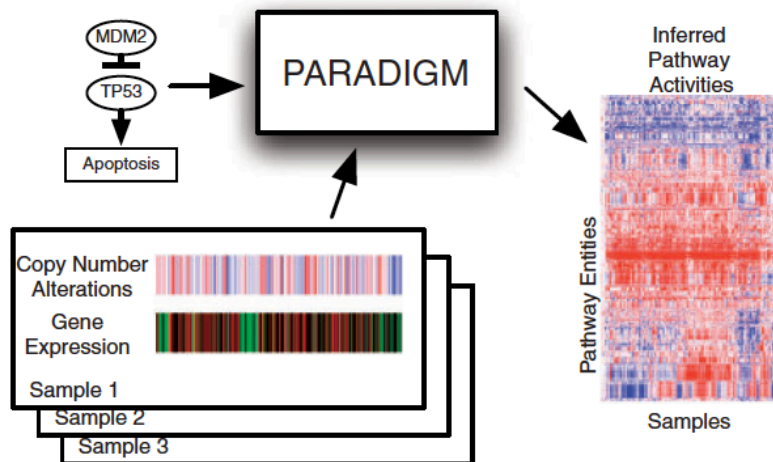
**Results:** We present a novel method for inferring patient-specific genetic activities incorporating curated pathway interactions among genes. A gene is modeled by a factor graph as a set of interconnected variables encoding the expression and known activity of a gene and its products, allowing the incorporation of many types of omic data as evidence. The method predicts the degree to which a pathway's activities (e.g. internal gene states, interactions or high-level 'outputs') are altered in the patient using probabilistic inference.

Compared with a competing pathway activity inference approach called SPIA, our method identifies altered activities in cancer-related pathways with fewer false-positives in both a glioblastoma multiform (GBM) and a breast cancer dataset. PARADIGM identified consistent pathway-level activities for subsets of the GBM patients that are overlooked when genes are considered in isolation. Further, grouping GBM patients based on their significant pathway perturbations divides them into clinically-relevant subgroups having significantly different survival outcomes. These findings suggest that therapeutics might be chosen that target genes at critical points in the commonly perturbed pathway(s) of a group of patients.

**Availability:** Source code available at <http://sbenz.github.com/Paradigm>

**Contact:** [jstuart@soe.ucsc.edu](mailto:jstuart@soe.ucsc.edu)

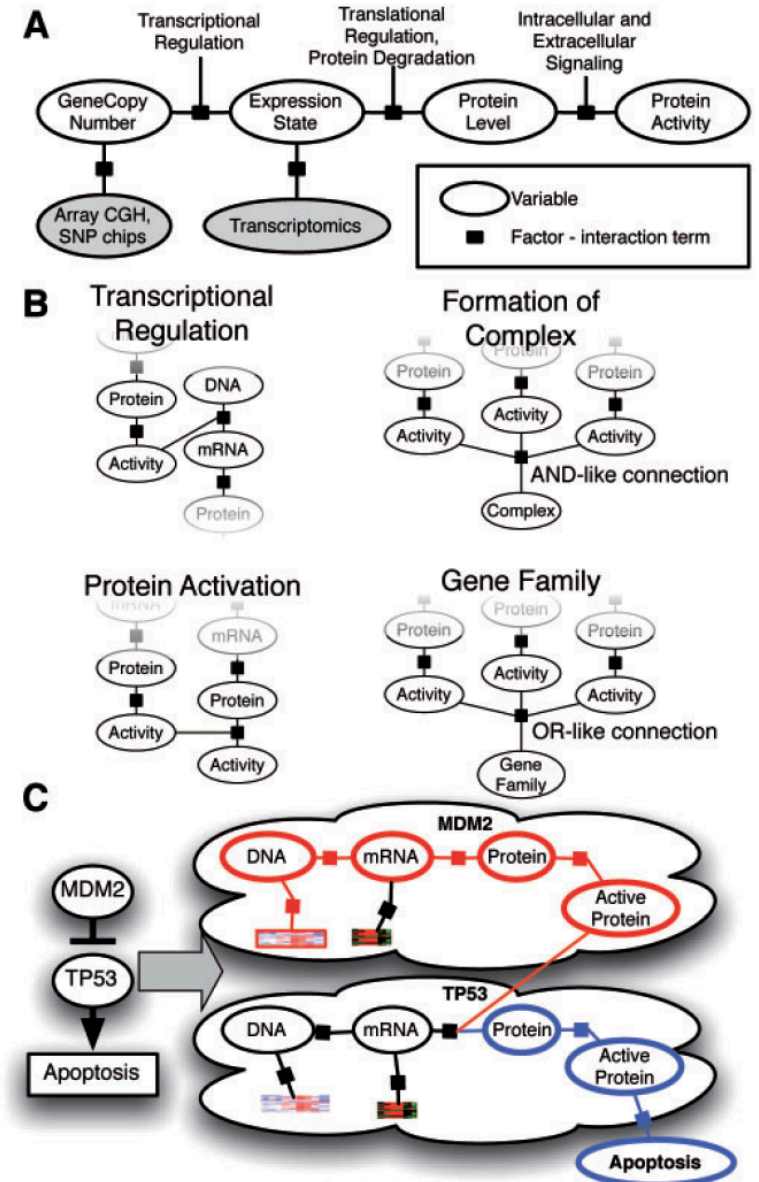
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.



**Fig. 2.** Overview of the PARADIGM method. PARADIGM uses a pathway schematic with functional genomic data to infer genetic activities that can be used for further downstream analysis.

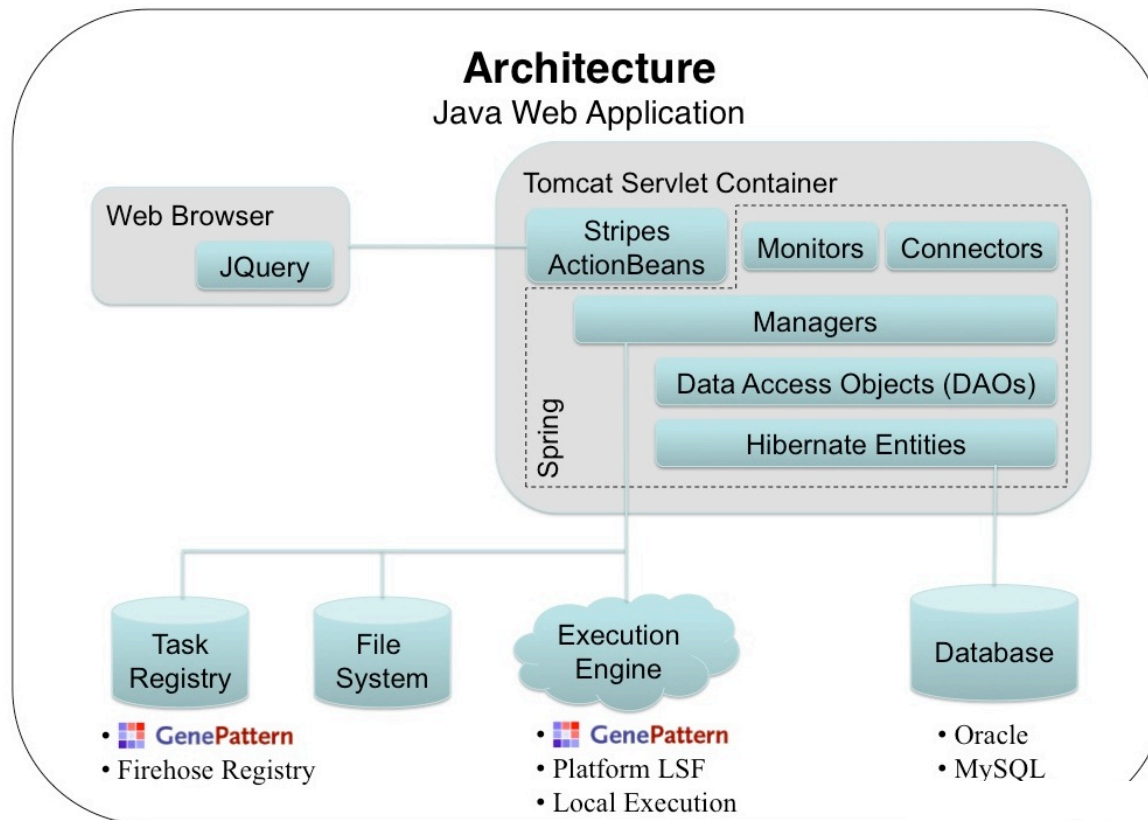
# PARADIGM

- Statistical Process that Can Incorporate
  - Protein Levels
  - Protein Activity levels
  - mRNA Transcript Levels
  - Copy-number levels for overall DNA
  - Actual Protein Pathway Interactions
- Use Factor graph to represent the biological pathway
- To Create Integrated Pathway Level (IPL) data for provided samples to be used in clustering and GSEA analysis
- Converting individual “omics” data for individual patient into biological pathway level activity.



# FIREHOSE

- Firehose is an analysis infrastructure developed at The Broad Institute to coordinate the flow of terabyte-scale datasets through dozens of quantitative algorithms.
- Although still evolving, Firehose has become a valuable piece of The Broad Institute computing infrastructure; it is used daily by dozens in the Cancer Genome Analysis group, to perform all TCGA GDAC and GSC analyses, managing hundreds of thousands of jobs on tens of thousands of samples, spread over hundreds of compute nodes and a 400+ TB file system.



# NOZZLE

## Nozzle: a report generation toolkit for data analysis pipelines

Nils Gehlenborg<sup>1,2</sup>, Michael S. Noble<sup>2</sup>, Gad Getz<sup>2</sup>, Lynda Chin<sup>2,3</sup> and Peter J. Park<sup>1,\*</sup>

<sup>1</sup>Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA, <sup>2</sup>Cancer Program, Broad Institute, Cambridge, MA 02142, USA and <sup>3</sup>Department of Genomic Medicine, MD Anderson Cancer Center, Houston, TX 77230, USA

Associate Editor: Martin Bishop

### ABSTRACT

**Summary:** We have developed Nozzle, an R package that provides an Application Programming Interface to generate HTML reports with dynamic user interface elements. Nozzle was designed to facilitate summarization and rapid browsing of complex results in data analysis pipelines where multiple analyses are performed frequently on big datasets. The package can be applied to any project where user-friendly reports need to be created.

**Availability:** The R package is available on CRAN at <http://cran.r-project.org/package=Nozzle.R1>. Examples and additional materials are available at <http://gdac.broadinstitute.org/nozzle>. The source code is also available at <http://www.github.com/parklab/Nozzle>.

**Contact:** [peter\\_park@hms.harvard.edu](mailto:peter_park@hms.harvard.edu)

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

### Phase 1: create report elements

```
r <- newCustomReport( "My Report" );
s <- newSection( "My Section" );
ss1 <- newSection( "My Subsection 1" );
ss2 <- newSection( "My Subsection 2" );
t <- newTable( iris[45:55,], "Iris data." );
p <- newParagraph( "Some sample text." );
```

### Phase 2: assemble report structure bottom-up

```
ss1 <- addTo( ss1, t ); # parent, child_1, ..., child_n
ss2 <- addTo( ss2, p );
s <- addTo( s, ss1, ss2 );
r <- addTo( r, s );
```

### Phase 3: render report to file

```
writeReport( r, filename="my_report" ); # w/o extension
```

Fig. 1. Sample R script to create a basic Nozzle report that illustrates the three phases of the bottom-up approach. See Supplementary Figure S1 for the HTML report

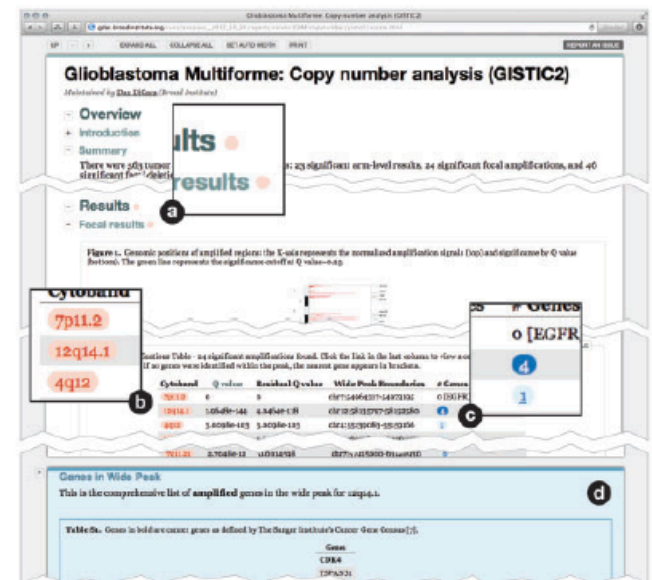


Fig. 2. A sample Nozzle report. (a) Red markers indicate statistically significant—as defined by the report author—results in this section. (b) Red boxes indicate significant results. (c) Underlined results have associated supplementary information. Clicking opens the (d) Supplementary Information panel

# UCSC Cancer Genome Browser

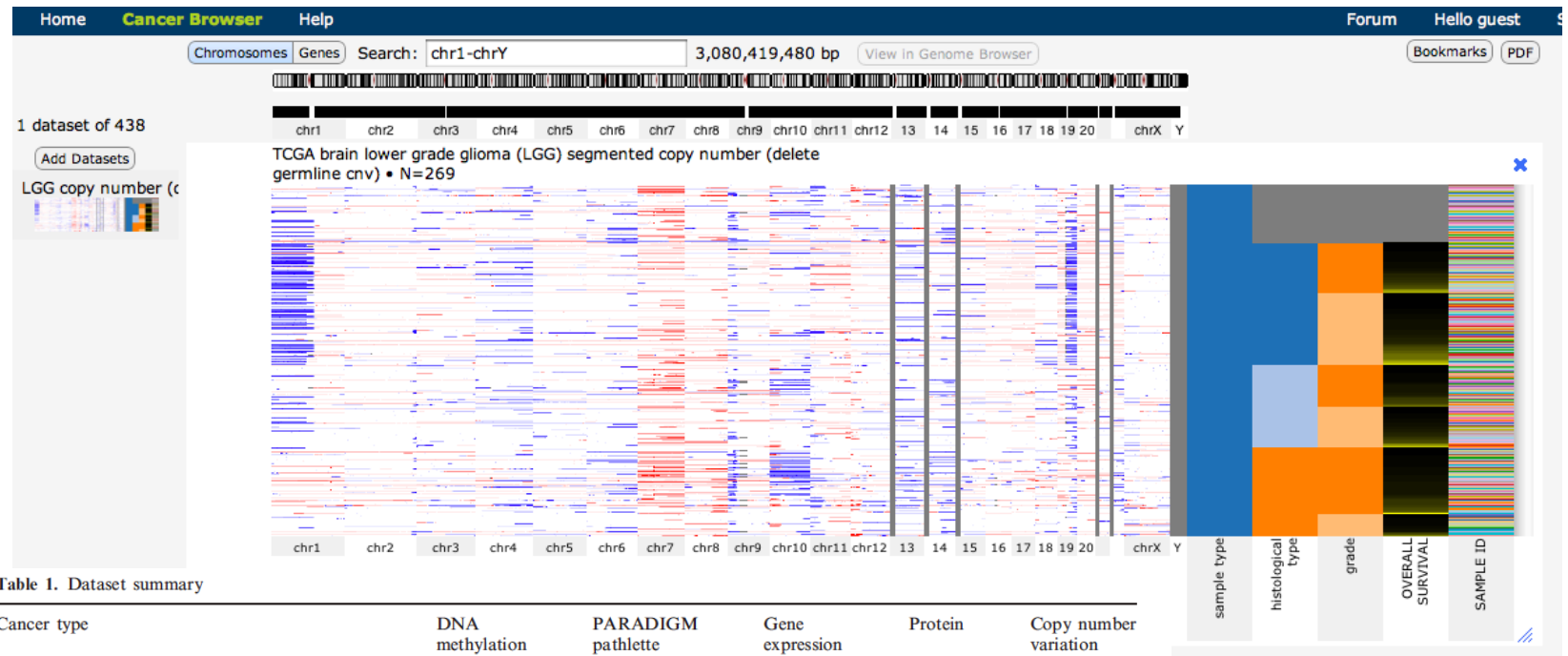
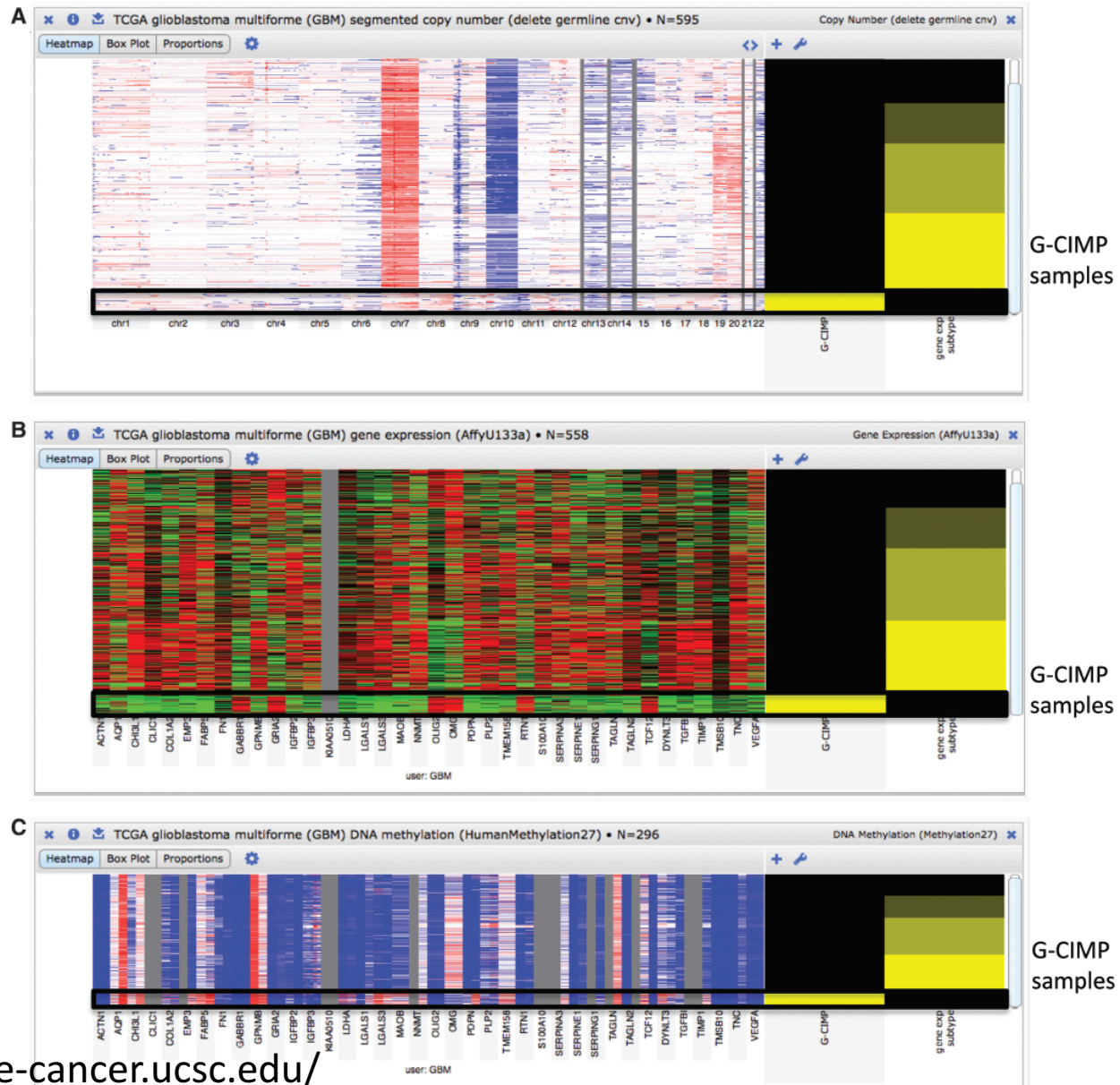


Table 1. Dataset summary

Cancer type	DNA methylation	PARADIGM pathlette	Gene expression	Protein	Copy number variation
TCGA acute myeloid leukemia	2 (384)		1 (179)		
TCGA bladder urothelial carcinoma	1 (91)		2 (106)		5 (221)
TCGA brain lower grade glioma	1 (26)	1 (27)	2 (54)		6 (608)
TCGA breast invasive carcinoma	2 (994)	1 (502)	4 (2870)	1 (410)	6 (4808)
TCGA cervical and endocervical SCC					5 (244)
TCGA colon and rectum adenocarcinoma	1 (236)	1 (208)	1 (224)	1 (463)	4 (2256)
TCGA colon adenocarcinoma	2 (498)		2 (366)		2 (894)
TCGA glioblastoma multiforme	2 (370)	1 (484)	4 (1693)	1 (215)	6 (3338)
TCGA head and neck squamous cell carcinoma	1 (342)		2 (555)		5 (1075)
TCGA kidney renal clear cell carcinoma	2 (861)	1 (69)	4 (1222)	1 (454)	6 (3028)
TCGA kidney renal papillary cell carcinoma	2 (146)	1 (16)	4 (126)		6 (340)
TCGA liver hepatocellular carcinoma			2 (52)		5 (275)
TCGA lung adenocarcinoma	2 (406)	1 (32)	4 (488)		6 (1374)
TCGA lung squamous cell carcinoma	2 (354)	1 (136)	5 (925)		6 (1442)
TCGA ovarian serous cystadenocarcinoma	1 (590)	1 (546)	4 (1761)	1 (412)	6 (3366)
TCGA pancreatic adenocarcinoma	1 (36)				5 (70)
TCGA prostate adenocarcinoma	1 (202)		1 (60)		5 (446)
TCGA rectum adenocarcinoma	2 (178)		2 (144)		2 (334)
TCGA skin cutaneous melanoma	1 (242)		1 (154)		2 (442)
TCGA stomach adenocarcinoma	2 (212)		1 (58)		5 (696)
TCGA thyroid carcinoma	1 (260)		1 (86)		5 (715)
TCGA uterine corpus endometrioid carcinoma	2 (488)	1 (53)	3 (448)	1 (200)	6 (2312)
SU2C Breast Public			1 (54)		2 (92)
CCLC			2 (1934)		1 (972)
Other datasets from the literature			19 (3556)		17 (2206)

Number of datasets by cancer type and data type; values in parenthesis are number of samples.

# UCSC Cancer Genome Browser



# NCI GDC Data Portal

## Harmonized Cancer Datasets

## Genomic Data Commons Data Portal

Get Started by Exploring:

Projects

Exploration

Analysis

Repository

Q e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

### Data Portal Summary

[Data Release 13.0 - September 27, 2018](#)

PROJECTS

43

PRIMARY SITES

69

CASES

33,096

FILES

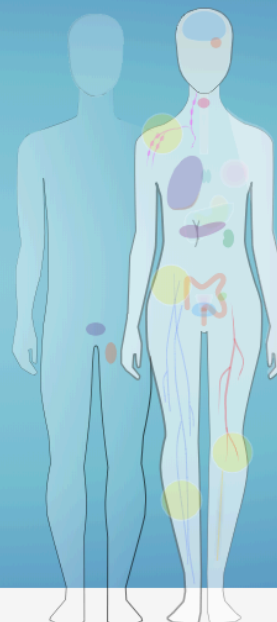
358,092

GENES

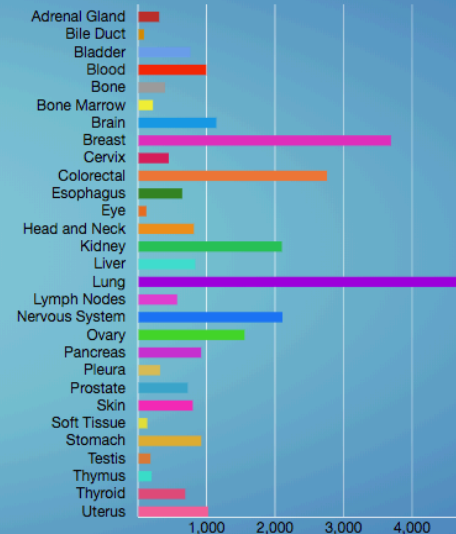
22,147

MUTATIONS

3,142,246



Cases by Major Primary Site



### GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:



Data Portal



Website



Data Transfer Tool



API



Data Submission Portal




Documentation



Legacy Archive

# ICGC Data Portal



## ICGC Data Portal

[Cancer Projects](#) [Advanced Search](#) [Data Repository](#)

Search

### About Us

The [ICGC Data Portal](#) provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.


For release 13 and earlier please see the [Legacy portal](#).


New features will be regularly added by the DCC development team. [Feedback](#) is welcome.

Subscribe to our Twitter [feed](#) to get updates.

### Tweets

[Follow @icgc\\_dcc](#)

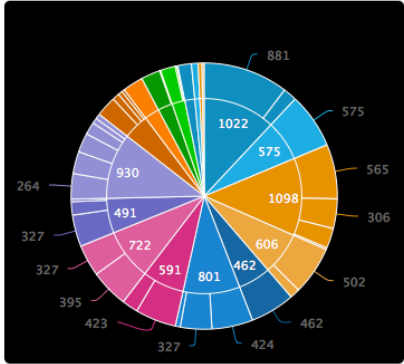
 **ICGC DCC** @icgc\_dcc  
OICR has been announced as a finalist for Cloudera's Data Impact Awards! See [goo.gl/P6UII3](http://goo.gl/P6UII3) for details!  
Expand

 **Jason H. Moore, Ph.D** @moorejrh  
My 2012 review of #GWAS with @vubush in #PLoS C has had over 21,000 views. [ploscompbiol.org/article/info%3A...](http://ploscompbiol.org/article/info%3A...) #genomics #bioinf  
Retweeted by ICGC DCC  
[Show Summary](#)

[See more news.](#)

### Data Release 14

September 26th, 2013



Cancer projects	41
Cancer primary sites	18
Donors	8,532
Simple somatic mutations	2,184,526
Mutated genes	54,682

### Information

[Access Raw Data](#)  
[Methods](#)  
[Submitter Tools](#)


### Tutorial

EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

[WATCH THE VIDEO TUTORIAL.](#)

SITE	INFORMATION	DCC	ICGC
<a href="#">Home</a>	<a href="#">Access Raw Data</a>	<a href="#">The Team</a>	<a href="#">ICGC Home</a>
<a href="#">Cancer Projects</a>	<a href="#">Methods</a>	<a href="#">Contact Us</a>	<a href="#">Data Access</a>
<a href="#">Advanced Search</a>	<a href="#">Submitter Tools</a>	<a href="#">Twitter</a>	<a href="#">Publication Policy</a>
<a href="#">Data Repository</a>		<a href="#">Legacy Portal</a>	<a href="#">Privacy Policy</a>
			<a href="#">Terms and Conditions</a>



© 2013 International Cancer Genome Consortium. All Rights reserved.



# cBio Portal for Cancer Genomics



Data Sets Web API R/MATLAB Tutorials FAQ News Visualize Your Data About

The cBioPortal for Cancer Genomics provides **visualization, analysis and download** of large-scale **cancer genomics** data sets. Please cite [Gao et al. \*Sci. Signal.\* 2013](#) & [Cerami et al. \*Cancer Discov.\* 2012](#) when publishing results based on cBioPortal.

QUERY DOWNLOAD DATA

**Select Studies:** 0 studies selected (0 samples) [Select all](#) Search...

Category	Count	Study Name	Sample Count
PanCancer Studies	2	MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017)	10945 samples
Cell lines	2	Pan-Lung Cancer (TCGA, Nat Genet 2016)	1144 samples
Adrenal Gland	1	Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012)	1019 samples
Ampulla of Vater	1	NCI-60 Cell Lines (NCI, Cancer Res. 2012)	60 samples
Biliary Tract	5	Adrenocortical Carcinoma (TCGA, Provisional)	92 samples
Bladder/Urinary Tract	7	Ampullary Carcinoma (Baylor College of Medicine, Cell Reports 2016)	160 samples
Blood	8		
Bone	2		
Bowel	5		
Breast	10		

**Select Data Type Priority:**  Mutation and CNA  Only Mutation  Only CNA

**Enter Gene Set:**    
Advanced: [Onco Query Language \(OQL\)](#)

**Submit Query** Please select one or more cancer studies.

## What's New

@cbioportal

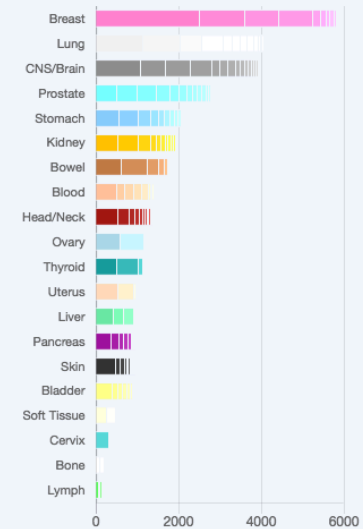
Sign up for low-volume email news alerts:

Subscribe

## Cancer Studies

The portal contains 166 cancer studies ([details](#))

### Cases by Top 20 Primary Sites



## Example Queries

- RAS/RAF alterations in colorectal cancer
- BRCA1 and BRCA2 mutations in ovarian cancer
- POLE hotspot mutations in endometrial cancer
- TP53 and MDM2/4 alterations in GBM
- PTEN mutations in GBM in text format
- BRAF V600E mutations across cancer types
- Patient view of an endometrial cancer case

## Testimonials

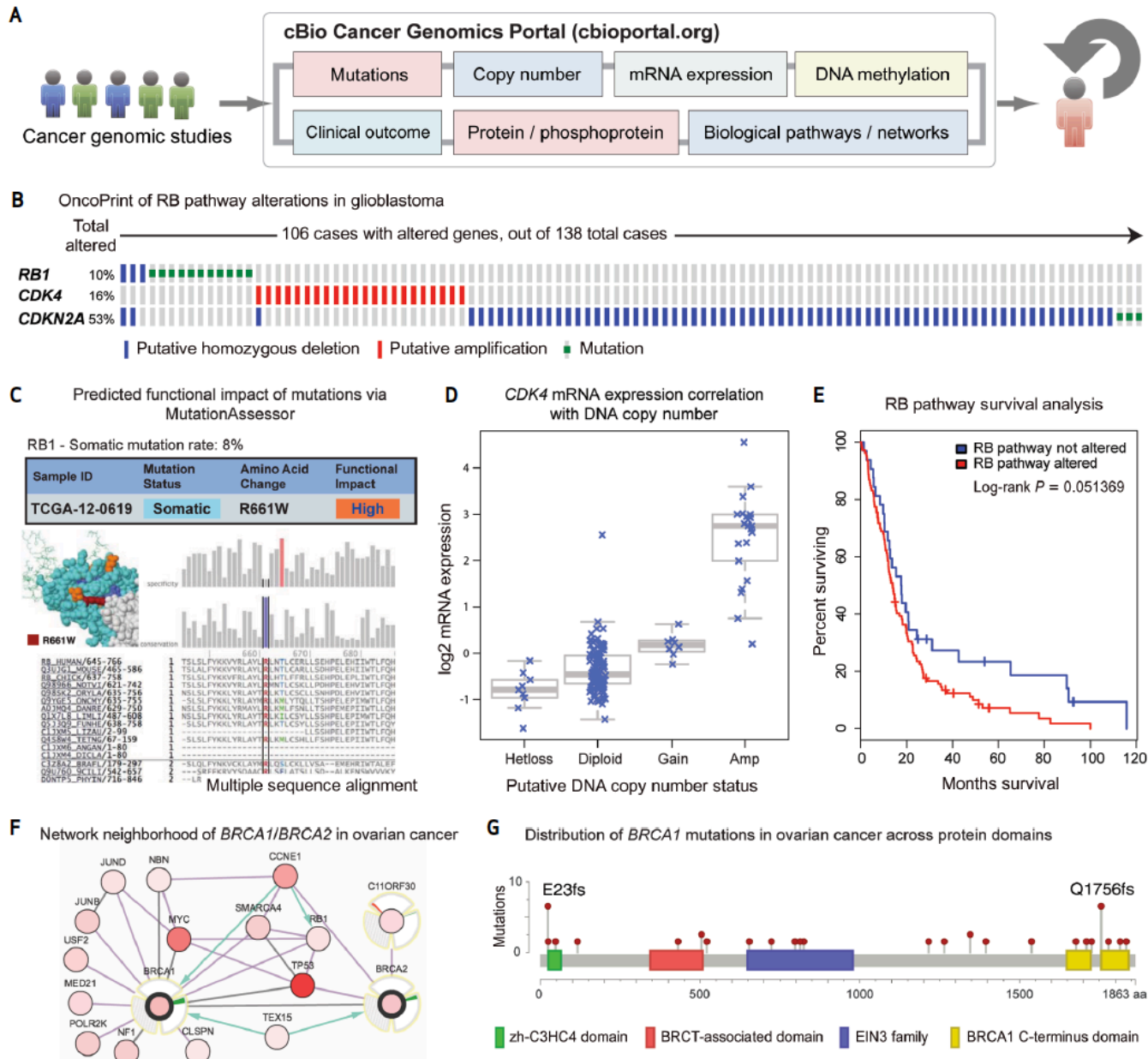
"I want to thank you for the nice, useful and user-friendly interface you have generated and shared with the community."

--Postdoctoral Fellow, Harvard Medical School, Children's Hospital Boston

[View All](#)

[Tell Us What You Think](#)

# cBio Portal for Cancer Genomics



# Take Home Message

---

- **Cancer is caused by genetic alterations:** Most human cancers are caused by two to eight sequential alterations that develop over the course of 20 to 30 years.
- **Comprehensive Characterization of Cancer Genomes:** can identify “driver” mutations that contribute to cancer tumorigenesis.
- **Drivers act in regulating core cellular processes:** The known driver genes function through a dozen signaling pathways that regulate three core cellular processes: cell fate determination, cell survival, and genome maintenance.
- **Pathway-centric Analysis of Cancer Genomes:** Every individual tumor, even of the same histopathologic subtype as another tumor, is distinct with respect to its genetic alterations, but the pathways affected in different tumors are similar.
- **Tumor Heterogeneity:** Genetic heterogeneity among the cells of an individual tumor always exists and can impact the response to therapeutics.
- **Bioinformatics Tools:** are needed to organize, analyze, interpret and visualize large-scale cancer genomics data.