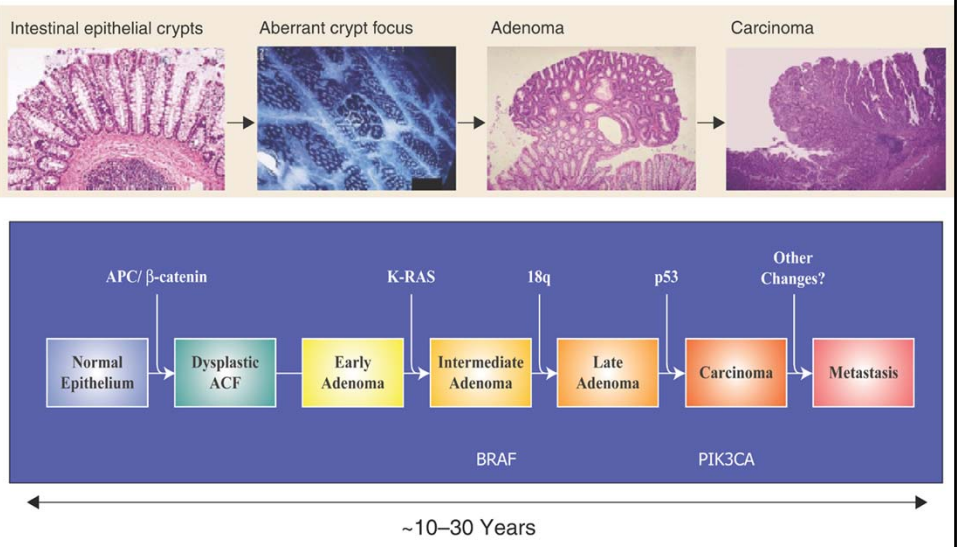
Whole Exome Sequencing to Identify Somatic Variants in Cancer

Yardena Samuels, Ph.D.
NHGRI / NIH

Next-Gen 101
A 'How to for Whole Exome Sequencing Research

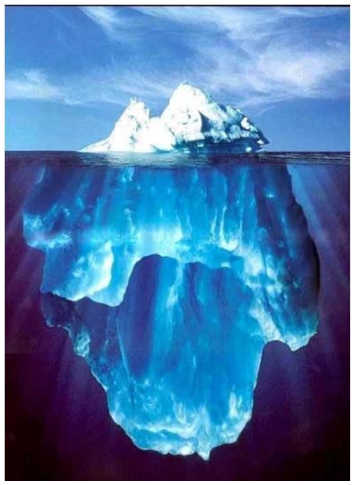
September 28th 2011

Cancer is a Genetic Disease



Cancer: Principles & Practice of Oncology, 9th edition

What We Know About Cancer Genetics



Overview

I. Platform setup for somatic mutation analysis of cancer genomes

II. Deciphering the cancer genetic landscape
   a. Genomic DNA source decisions
   b. Quality test of whole exome data
   c. Necessary data to evaluate 'drivers' and 'passengers'
   d. Complex exomes derived from fresh tumors

Hurdles of High Throughput Sequencing

I. Establishing a high quality tissue bank

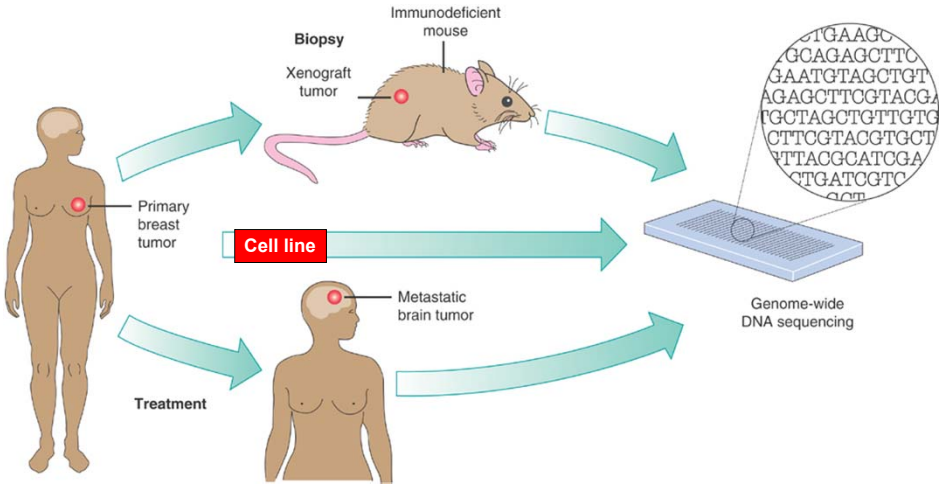
II. Sequencing large quantities of samples

III. Analyzing millions of bp to hunt for mutations

Initial Discoveries of Unbiased Sequencing Approaches

Gene	Cancer	Team
BRAF	Melanoma, thyroid, colorectal…	Sanger
PIK3CA	Colon, breast, liver, ….	Johns Hopkins

Tumor Bank Establishment-I



Slide adapted from Gray et al. Nature 2010

Tumor Bank Establishment-II

Tumor DNA source	Advantage/s	Challenge/s
Fresh frozen/OCT block	Highly reliable data	Limited DNA Heterogeneous Labor intensive extraction
Paraffin embedded tissue	Highly reliable data	Limited DNA Heterogeneous Labor intensive extraction DNA quality issues
Cell line	Plenty DNA Homogenous Simple extraction Functional studies	Genetic validation in fresh tumor
Xenograft	Plenty DNA Homogenous Simple extraction	Genetic validation in fresh tumor Expensive Mouse DNA contamination

Tumor Bank Establishment-III

- Normal tissue
  - -Blood-Not always available
  - -Neighboring tissue
    - Might have 'contaminating' tumor cells

- Clinical information
  - -DOB
  - -DOD
  - -Date of diagnosis
  - -Malignancy stage
  - -Location of primary tumor
  - -Location of metastatic tumor
  - -Therapies

Tumor Bank Establishment-IV

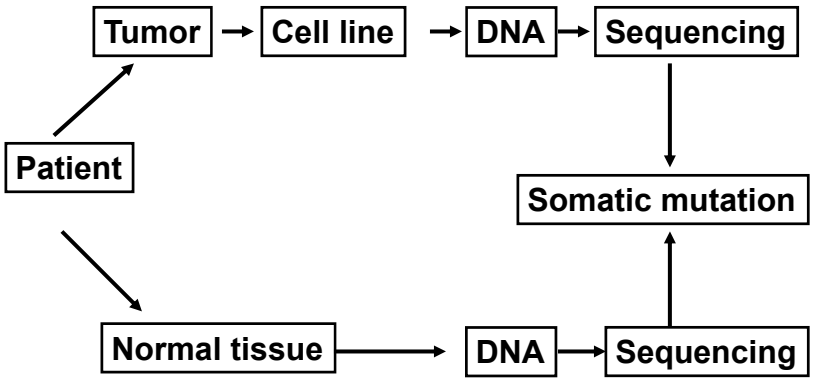
	Sample cohort #1	Sample cohort #2	Sample cohort #3
Metastatic tumor DNA	120	32	40
Matched normal DNA	Yes	Yes	Yes
OCT blocks	Yes	Yes	Yes
Matched cell line	Yes	No	No
Matched RNA	Yes	No	No
Matched protein lysate	Yes	No	No
Clinical Information	Yes	Yes	Yes

Importance in acquiring additional patient cohorts in order to validate the genetic data

Tumor Bank Quality Controls

- SNP detection to make sure the tumor and normal tissues are matched
- Implement an assay to determine that the fraction of tumor cells is > 75%
- Mutational analysis of highly mutated genes in melanoma

Somatic Mutation Analysis



Methods of Mutation Hunting

Candidate approach

Whole Exome/Genome

The Cancer Genome Atlas (TCGA)

Launched in 2006 as a pilot and expanded in 2009.
The goal of TCGA is:

To provide comprehensive genomic characterization and sequencing data to the research community on at least 3,000 new cancer cases by the end of September 2011.

Slide adapted from Kenna Shaw, Ph.D. Deputy Director, TCGA Program Office , NCI
Brad Ozenberger, Ph.D. Program Director, NHGRI

Overview

I. Platform setup for somatic mutation analysis of cancer genomes

II. Deciphering the cancer genetic landscape
   a. Genomic DNA source decisions
   b. Quality test of whole exome data
   c. Necessary data to evaluate 'drivers' and 'passengers'
   d. Complex exomes derived from fresh tumors

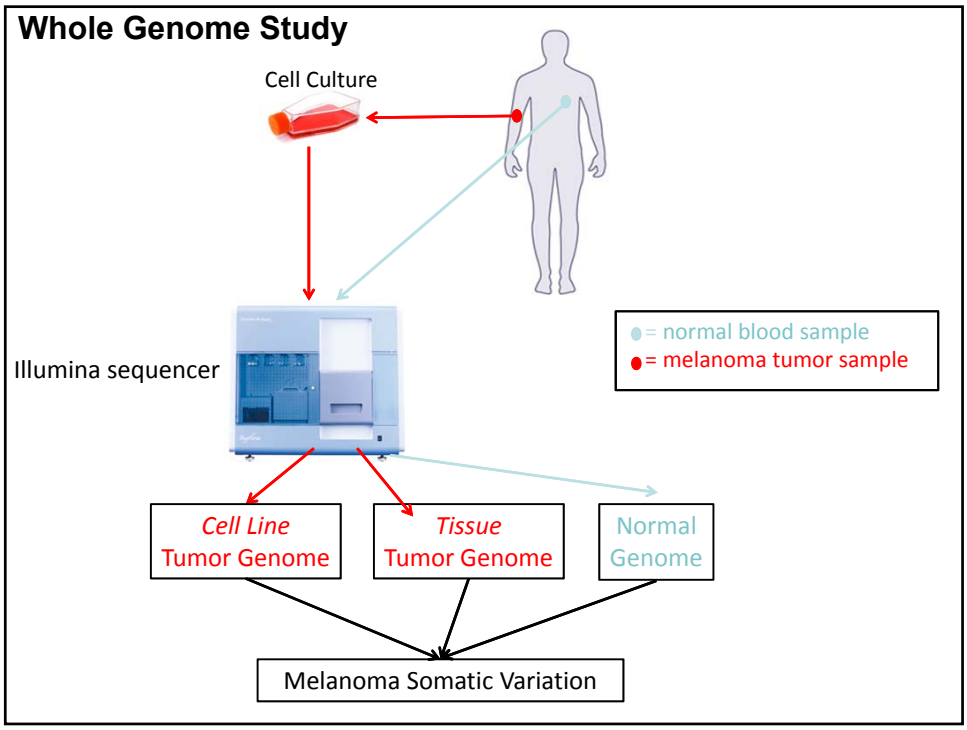
Whole Exome DNA Source

Fresh tumor?
Low passage cell line?

	Fresh Tumor	Cell line
DNA quantity	Limited DNA	Unlimited
Homogeneity/heterogeneity	Heterogeneous	Homogenous
Recapitulates tumor biology	Yes	??

How to choose DNA source for exome sequencing?

Whole genome sequencing as tool to assess

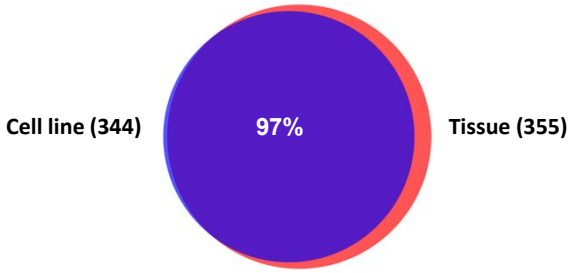


Build Statistics

	Tumor Cell Line	Tumor Tissue	Merged Normal
Read length	**2 x 100 bases**	**2 x 100 bases**	**2 x 100 bases**
Passing filter depth of coverage	**34x**	**37x**	**67x**

Aim to get 92% callable genotypes across the entire genome

Intersection of Fresh Tumor and Matched Cell Line



Intersection of non-synonymous and nonsense somatic variants in CDS

However, copy number variations were less concordant:
78.9% of tissue CNVs overlap with cell culture CNVs

Whole Exome DNA Source

Fresh tumor?
Low passage cell line?

We used low passage cell line derived genomic DNA as:

-The SNV data will be concordant with fresh tumor SNVs

-Whole exome capture required large amounts of DNA (6 μgs)

-There will be no stroma "contamination"

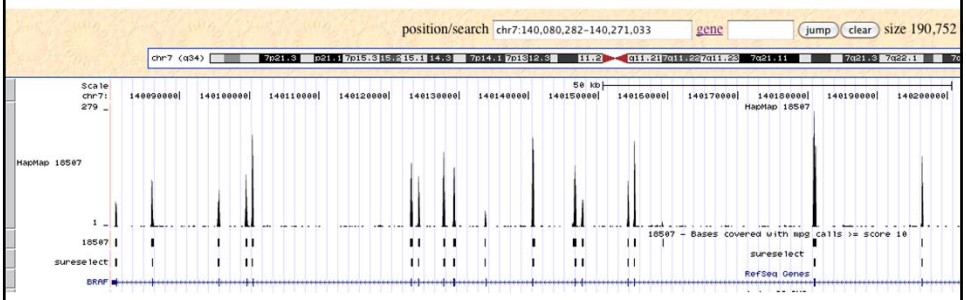
Whole Exome Sequencing Study Design

Discovery Exome capture (14 tumors/ matched normal)
Agilent SureSelect 37Mb
~20,000 genes and flanking regions
Illumina GAII platform
ELAND followed by cross_match

Validation Sanger

Wei et al, *Nature Genetics*, [Epub ahead of print] (2011)

SureSelect Capture of BRAF



Overview

I. Platform setup for somatic mutation analysis of cancer genomes

II. Deciphering the cancer genetic landscape

a. Genomic DNA source decisions

b. Quality test of whole exome data

c. Necessary data to evaluate ‘drivers’ and ‘passengers’

d. Complex exomes derived from fresh tumors

Quality Tests of Whole Exome Data

1. **Target region genotype coverage**
2. Specificity assessment
3. Sensitivity assessment
4. Number of somatic mutations per tumor
5. Potential artifacts

1 Target Region Genotype Coverage

Sample	Fold coverage over baited exome	% target region genotype coverage*
01N	259	90
01T	278	86
05N	187	86
05T	184	87
09N	278	89
09T	272	86
12N	339	91
12T	336	91
18N	208	93
18T	257	92
22N	209	90
22T	276	89

1 Whole Exome Sequencing Performance

- ~ 12 Gb of sequence per sample
- Depth >180X
- Exome with >90% covered by high quality genotypes

Quality Tests of Whole Exome Data

1. Target region genotype coverage
2. **Specificity assessment**
3. Sensitivity assessment
4. Number of somatic mutations per tumor
5. Potential artifacts

2 Specificity Assessment

Whole exome score cutoff for determination of somatic mutations
Positives

Refseq	Ref_allele	Var_allele	Ref_aa	Var_aa	Normal name	Normal MPG/coverage	Tumor name	Tumor MPG/coverage	Sanger evaluation
DNAH5	C	T	E	K	55N	0.70	55T	0.50	somatic mutation
CHL1	C	T	H	Y	24N	70.00	24T	0.54	somatic mutation
NOS1	G	A	S	L	24N	74.00	24T	0.63	somatic mutation
DCC	G	A	G	E	12N	35.00	12T	0.66	somatic mutation
BRAF	A	T	V	E	22N	0.71	22T	0.68	somatic mutation

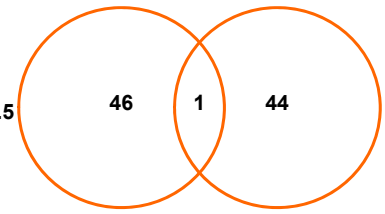
Whole exome score cutoff for determination of somatic mutations
Negatives

Refseq	Ref_allele	Var_allele	Ref_aa	Var_aa	Normal name	Normal MPG/coverage	Tumor name	Tumor MPG/coverage	Sanger evaluation
RBMX	A	G	L	P	91N	0.08	91T	0.33	no mutation
EEF1B2	T	C	S	G	51N	0.11	51T	0.43	no mutation
ARHGAP21	G	C	N	K	12N	0.12	12T	0.52	no mutation
PABPC1	G	A	S	L	5N	0.13	5T	0.48	no mutation
AP3S1	A	G	N	S	22N	0.15	22T	0.23	no mutation
AP3S1	A	G	N	S	96N	0.16	96T	0.04	no mutation

2 Specificity Assessment

91 regions assessed by Sanger sequencing

Validated
MPG/coverage ratio >0.5



Non-Validated
MPG/coverage ratio <0.5

- **97.9%** coverage rate
- **2.4%** false negative rate
- **18%** of the alterations removed

MPG= **M**ost **P**robable **G**enotype. Use MPG >= 10

Quality Tests of Whole Exome Data

1. Target region genotype coverage
2. Specificity assessment
3. **Sensitivity assessment**
4. Number of somatic mutations per tumor
5. Potential artifacts

3 Sensitivity Assessment

Out of 47 somatic substitutions discovered by candidate approach

38 were present in our whole exome study.

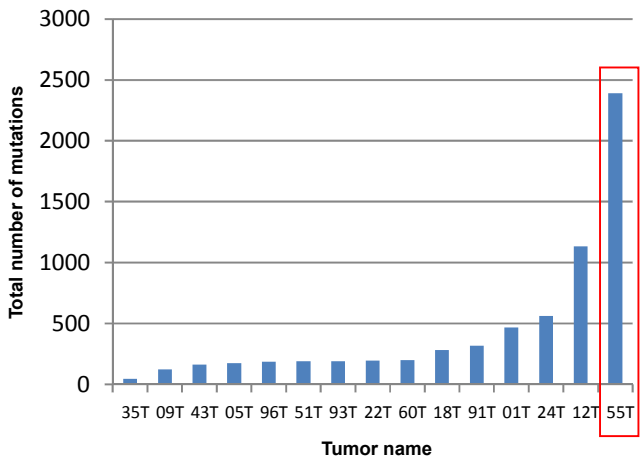
81% sensitivity

Note-the missed alterations were captured and well covered-simply missed by the exome study.

Quality Tests of Whole Exome Data

1. Target region genotype coverage
2. Specificity assessment
3. Sensitivity assessment
4. **Number of somatic mutations per tumor**
5. Potential artifacts

4 Number of Somatic Mutations Per Tumor



Quality Tests of Whole Exome Data

1. Target region genotype coverage
2. Specificity assessment
3. Sensitivity assessment
4. Number of somatic mutations per tumor
5. **Potential artifacts**

5 Potential Artifacts Due to Chromosome Duplication

When looking at the data it is important to sort it not only by sample, but also by chromosome.

When this was done for patient 9, there seemed to be an out of the ordinary number of somatic mutations on chromosome X.

So we looked more closely at this and found that---

The genotypes on Chr:X in 9N had <u>one allele</u>, -> patient is male
However, his tumor had <u>two alleles</u> in the same precise location

Chr	LeftFlank	RightFlank	refseq	transcript	type	09N_norm_control.NA	09T_aff_case.NA
chrX	3239493	3239495	MXRA5	uc004crg.2	Non-synonymous	C	CC
chrX	3540332	3540334	PRKX	uc010nde.1	Non-synonymous	C	CC
chrX	3543871	3543873	PRKX	uc010nde.1	Non-synonymous	G	GG
chrX	5821090	5821092	NLGN4X	uc010ndj.1	Non-synonymous	C	CC
chrX	6461790	6461792	VCX3A	uc004crs.1	Non-synonymous	C	CC
chrX	6461808	6461810	VCX3A	uc004crs.1	Non-synonymous	C	CC
chrX	7771780	7771782	VCX2	uc010ndn.1	Non-synonymous	T	TT
chrX	7771910	7771912	VCX	uc004crz.1	Non-synonymous	G	GG
chrX	8394143	8394145	VCX3A	uc004cse.1	Non-synonymous	T	TT

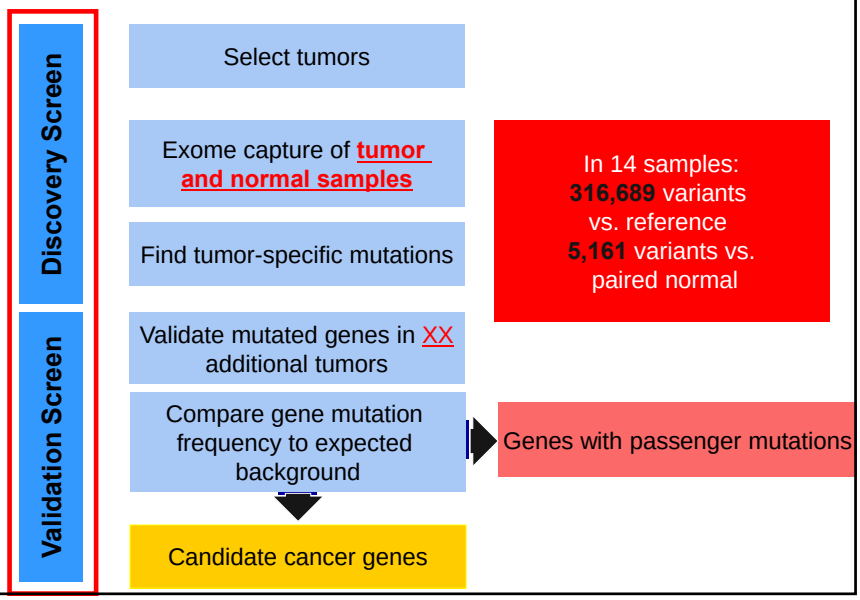
Thus, copy number variation occurred:
Y chromosome deletion vs.
X chromosome duplication

Need to investigate the underlying reason before including these alterations in chromosome X in patient 9.

Quality Tests of Whole Exome Data

1. **Target region genotype coverage**
2. **Specificity assessment**
3. **Sensitivity assessment**
4. **Number of somatic mutations per tumor**
5. **Potential artifacts**

Exome-Wide Mutational Analyses Study Design



Insights Into the Filtering Process

Number of potential somatic mutations: **316,689**

↓

Filter through **dbSNP**: **141,961**

↓

Apply **somatic** filter: **58,352**

↓

Some mutations appear somatic in one sample but appear in another normal sample-these are removed: **19,960**

↓

Remove all non-coding alterations: **5,161**

Whole Exome Discovery Screen

Exome capture sequencing of **14** untreated melanoma samples and their matched normal
180x coverage, 90% bases with high quality genotype calls

↓

Align sequence data (genome build hg18) and filter putative somatic mutations

↓

Number of potential somatic mutations: **5,161**

↓

Number of mutations with a MPG/Coverage ratio>=0.5: **4,226**

↓

Missense/ nonsense/ splice site mutations: **2,813**
Insertions/Deletions: **27**
Synonymous mutations: **1,386**

Nonsynonymous:synonymous ratio 2:1

Overview

I. Platform setup for somatic mutation analysis of cancer genomes

II. Deciphering the cancer genetic landscape
   a. Genomic DNA source decisions
   b. Quality test of whole exome data
   c. Necessary data to evaluate 'drivers' and 'passengers'
   d. Complex exomes derived from fresh tumors

The Challenge in Cancer Genomics
‘Passengers’ versus ‘Drivers’

- **Statistics**
- **Bioinformatics**
- **Functional studies**

The Challenge in Cancer Genomics
‘Passengers’ versus ‘Drivers’

- **Statistics**

1 **- Nonsynonymous: synonymous ratio**

Nonsynonymous:synonymous ratio
2:1

2 **- Mutations above background mutation rate**

The background mutation rate is the number of mutations per megabase DNA derived from all your exomes.

In melanoma the background mutation rate is 11.4 mut/Mb

The Challenge in Cancer Genomics
'Passengers' versus 'Drivers'

- **Statistics**
  - 1 - Nonsynonymous: synonymous ratio
  - 2 - Mutations above background mutation rate
  - 3 - Recurrantly mutated genes: "Hotspots"
  - 4 - Highly mutated genes
- **Bioinformatics**
- **Functional studies**

Validation Screen

Search for recurrent "Hotspot" mutations

↓

9 novel genes with recurring mutations

↘

Validate mutated genes in Additional tumors

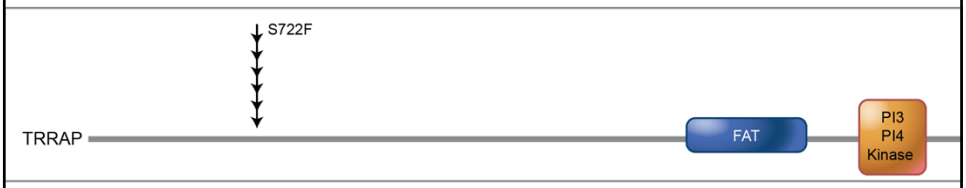
Our set { Discovery (n=14)
Prevalence (n=70)

Validation set 1 (n=39)
Validation set 2 (n=32)
Commercial cell lines (n=12)

Validated Recurrent Mutations

Gene Name	# of Tumors Affected	Nucleotide Change	Amino Acid Change	Synonymous or Nonsynonymous	Tumor Name	Tumor Panel
CPT1A	2	C1638T	F546F	Synonymous	5T	Exome Capture
					43T	Exome Capture
DCC	3	G164A	G55E	Nonsynonymous	12T	Exome Capture
					18T	Exome Capture
					MB1160_T	Validation set 1
FCRL1	3	C741T	I247I	Synonymous	91T	Exome Capture
					96T	Exome Capture
					63T	Prevalence screen
LRRN3	2	G1084A	E362K	Nonsynonymous	12T	Exome Capture
					24T	Exome Capture
NOS1	2	C2312T	S771L	Nonsynonymous	24T	Exome Capture
					60T	Exome Capture
PLCH1	2	C907T	Q303X	Nonsynonymous	1T	Exome Capture
					24T	Exome Capture
SLC17A5	2	C1090T	R364C	Nonsynonymous	12T	Exome Capture
					18T	Exome Capture
TRRAP	6	C2165T	S722F	Nonsynonymous	63T	Exome Capture
					91T	Exome Capture
					96T	Prevalence screen
					106T	Prevalence screen
					119T	Prevalence screen
					A375	Commercial cell line
ZNF831	3	C4421T	S1474F	Nonsynonymous	43T	Exome Capture
					91T	Exome Capture
					MB1160_T	Validation set 1

Distribution of novel nonsynonymous recurrent mutations



The likelihood for the occurrence of 6 identical mutations is approximately $5x10^{-20}$

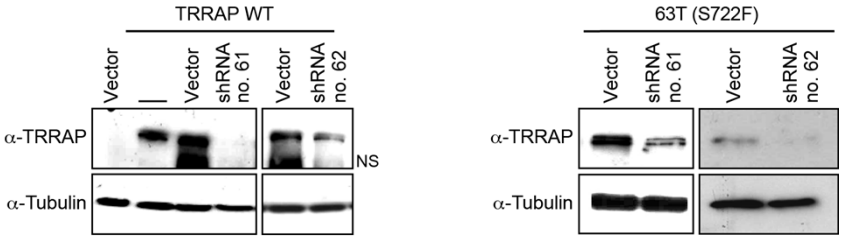
- Functions as part of a histone acetyltranferase complex
- Disruption of TRRAP causes defects in cell cycle progression

Comparison of Conserved Serine-722 of Human TRRAP with its Orthologs

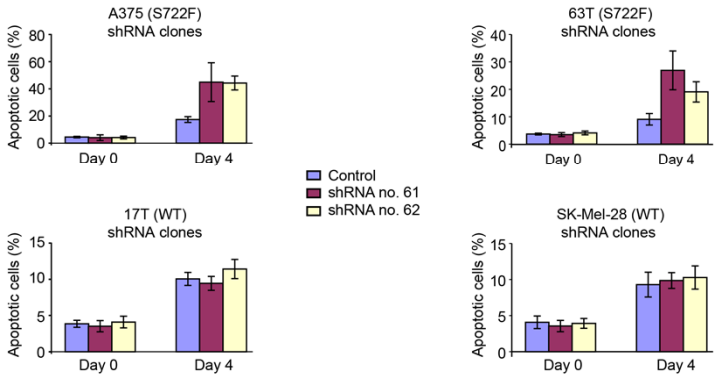
TRRAP S722F (C2164T)

```
NP_003487 Homo sapiens        697 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 747
XP_860949 Canis familiaris    696 LPEMGSHVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 746
XP_001136733 Pan troglodytes  697 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 747
XP_583735 Bos taurus          698 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 748
XP_213706 Rattus norvegicus   699 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 749
XP_414752 Gallus gallus       685 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 735
XP_001919276 Danio rerio      652 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 702
NP_001074831 Mus musculus     697 LPEMGSNVE--LSNLYLKLFKLVFGSVSLFAA--ENEQMLKPHLHKIVNSSMELA 747
NP_001097192 D. melanogaster  664 MEEMGSNLE--RSNLYLRLFKLVFGSVSLFPV--ENEQMLRPHLHKIVNRSMELA 704
XP_556172 Anopheles gambiae   708 MDEMGSNIE--RSNLYLRLFKLVFGSVSLFAA--ENEHMLRPHLHNIVNRSMELA 748
NP_001022032 C. elegans       723 MKLLEVSND--KTMLYVKLFKIIFSAIGANGSGLHGDKMLTSYLPEILKQSTVLA 775
NP_011967 S. cerevisiae       748 LKDLG-NVDFNTSNVLIRLFKLSFMSVNLFPN--INEVVLLPHLNDLILNSLKYS 799
```

Effect of Mutant TRRAP on Apoptosis



TRRAP Mutation Confers Resistance to Apoptosis



Validation Screen

Search for recurrent "hotspot" mutations

Search for highly mutated genes

9 novel genes with recurring mutations

4 **16** highly mutated genes (binomial p value < 0.05; mutated in > than 2 discovery samples

Validate mutated genes in Additional tumors

Discovery (n=14)
Prevalence (n=38)

4 **Accounts for:**
-Transcript size (always use the longest transcript)
-Background mutation rate

Whole Exome Sequencing
validation step done incorrectly

Gene Name	% of tumor affected
MUC17	50.0
GRIN2A	42.9
DNAH5	42.9
SCN1A	35.7
DNAH7	35.7
TTN	35.7
CCDC63	28.6
TMEM132B	28.6
ZNF831	28.6
PLCB4	28.6
SALL1	28.6
CREBBP	28.6
ASH1L	28.6
XIRP2	28.6
CSMD2	28.6
DNAH2	28.6

Whole Exome Sequencing
validation step done right

		Exome Capture (n=14)	Combined Exome Capture and Prevalence Screens (n=52)
Gene name	P value	% of tumors affected	% of tumors affected
BRAF	4.80E-05	50%	65%
GRIN2A	6.36E-03	43%	33%
CCDC63	3.34E-03	29%	11%
TMEM132B	7.59E-03	29%	17%
ZNF831	1.29E-02	29%	17%
PLCB4	4.39E-02	29%	15%
AKR1B10	5.21E-03	21%	8%
TAS2R60	5.46E-03	21%	9%
KHDRBS2	7.26E-03	21%	9%
PTPRO	9.09E-03	21%	8%
SYT4	1.23E-02	21%	8%
UGT2B10	2.13E-02	21%	8%
SLC6A11	2.84E-02	21%	6%
SLC17A5	7.91E-03	21%	6%
C12orf63	4.46E-02	21%	9%
PCDHB8	4.80E-02	21%	8%

Validation of *GRIN2A* Mutations in <u>Two additional Cohorts</u>

GRIN2A mutated in **33%** of discovery (n=14) and prevalence samples (n=38)

GRIN2A mutated in **28.2%** of validation set 1 samples (n=39)

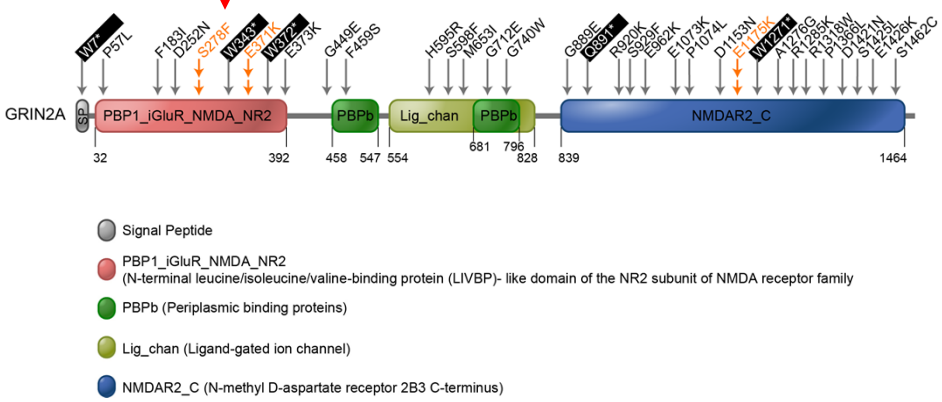
GRIN2A mutated in **15.6%** of validation set 2 samples (n=32)

GRIN2A mutated in **8.3%** of commercial cell lines (n=12)

Importance in acquiring additional patient cohorts in order to validate the genetic data

GRIN2A is Highly Mutated in Melanoma (33%)

Found in COSMIC



Overview

I. Platform setup for somatic mutation analysis of cancer genomes

II. Deciphering the cancer genetic landscape

a. Genomic DNA source decisions

b. Quality test of whole exome data

c. Necessary data to evaluate 'drivers' and 'passengers'

d. Complex exomes derived from fresh tumors

Whole Exome Derived From Fresh Tumors

Possible issues:

-Used of similar MPG and ratio criteria as used above

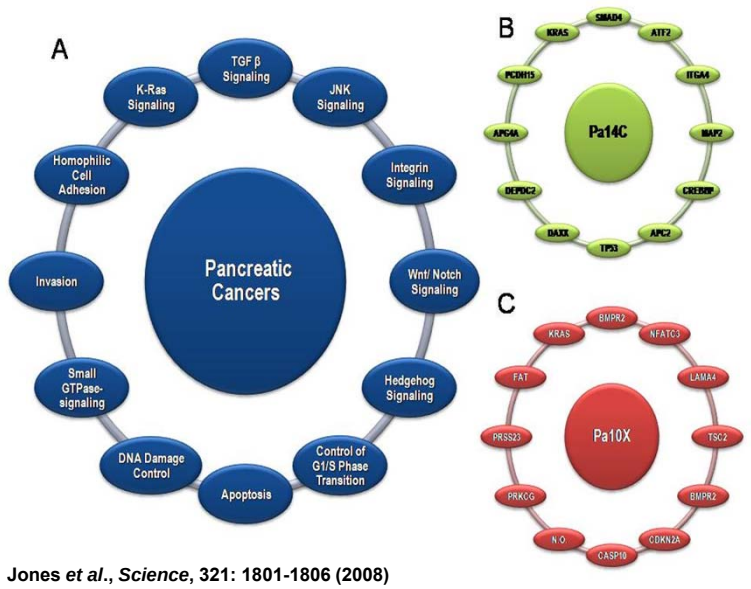
Find that somatic mutations identified in the tumor are also found in the normal sample (eg BRAF V600E)

Thus: tumor cells are "contaminating" the extracted normal tissue

-Heterogeneity-yet to be determined

Pathway Oriented Models in Cancer Genetics

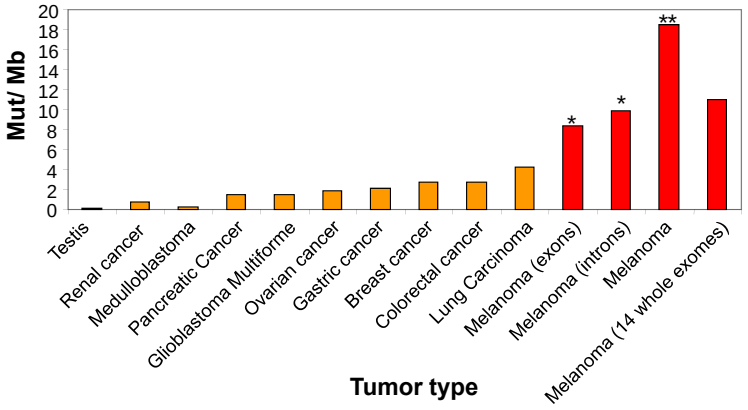
12 Core Pathways in Pancreatic Cancer



Jones *et al*., *Science*, 321: 1801-1806 (2008)

Delving Deeper into the Genome

Mutation Frequency in Solid Cancers



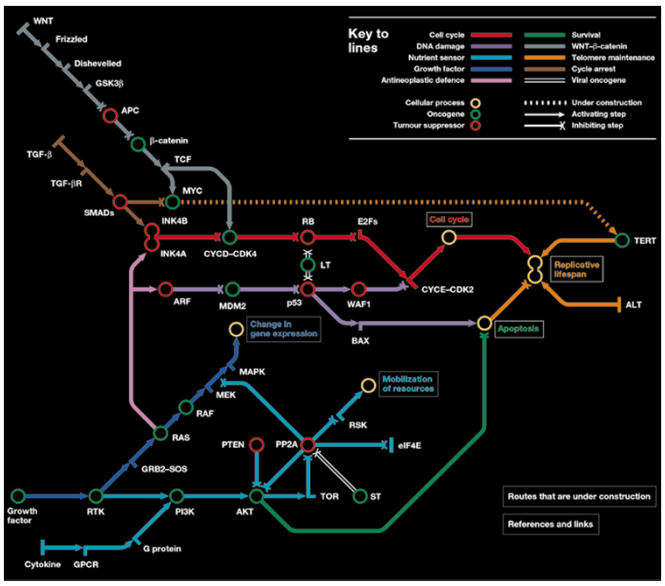
* Pleasance et al., Nature. 2010 Jan 14;463(7278):191-6

** Greenman et al., Nature. 2007 Mar 8;446(7132):153-8.

Future Challenges

- "Drivers" vs. "Passengers"
- How do we analyze and then interpret all the data?
- How do we perform high-throughput functional analysis?
- How do we apply the data to the clinic?

Mutational Database of Signal Transduction Pathways in Cancer



2 Specificity Assessment

These refinements gave us a 97.9% coverage rate, 2.4% false-negative rate

	Sanger Result	MPG score	# of samples
Total # of alterations tested using Sanger Sequencing: 91	47 Confirmed	≥0.5	46
		<0.5	1
	44 Not confirmed	≥0.5	4
		<0.5	40

Coverage study: 46/47=97.9%

False negative: $\frac{1}{40+1}$ = 2.4%