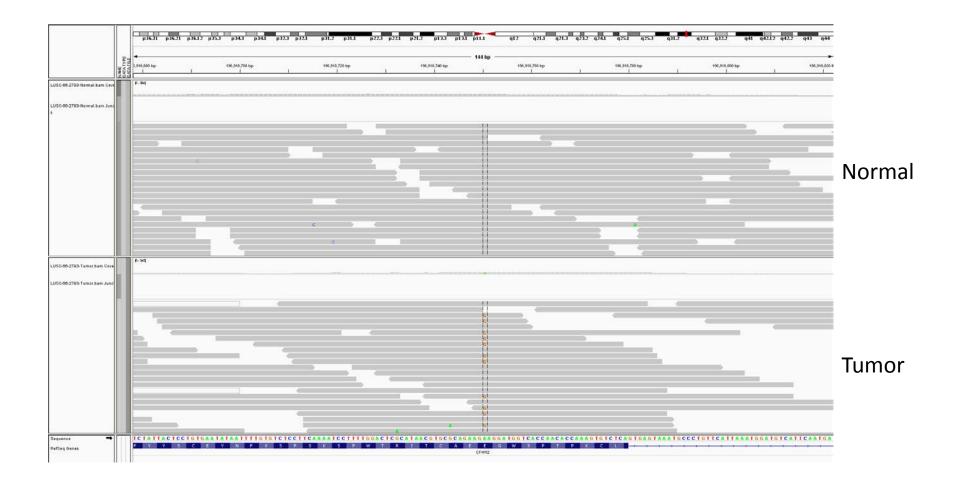
# Comparison and Validation of Somatic Mutation Callers

Andrey Sivachenko TCGA Symposium, November 17-18, 2011

## SNVs are Defined Simply...

- (single nucleotide) differences from the reference
- Ideally: resequence and read the results out
- If only everything looked as in the example below



### ... but SNVs Can be Hard to Call

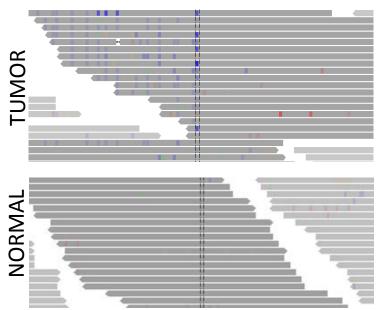
- Multiple issues in library preparation, sequencing and data processing (base calling, alignments) can result in a spectrum of SNV-like events, from good to terrible
- Need to watch for:
  - Alignment quality around the event
  - "Strandness" orientation of supporting reads
  - Position in read
  - Sufficient coverage (both in tumor *and* normal)
  - Sequence context
  - Potential tumor contamination in normal
  - p3631 p3621 p3612 p353 p343 p341 p323 p321 p22.3 p22.1 p21.2 p13.3 p13.1 p11.1 g12 g21.1 g21.3 g23.2 g24.1 g25.1 g25.3 12,897,740 b 12,897,760 b 12,897,790 ( 12.887.820 be 12,897,840 D SC-88-2703 Tumor ham C SC-66-2793 Tu teffing Gene PRAMEF1

### Specificity $\rightarrow$ Need to protect against two types of errors

Signal: ~1 somatic mutation per Mb Goal: >95% validation rate and ideally approach 100%

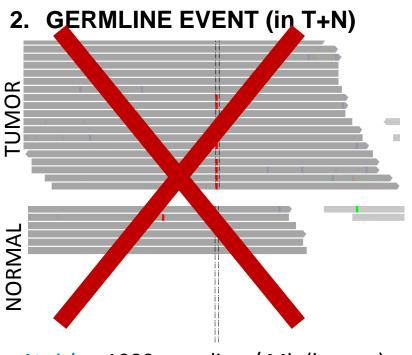
→ Need error rate to be  $\leq 0.05$  errors/Mb! 99.9999% is not good enough

Noise: Two types of false positives



1. NO EVENT

At risk: Every base Source: Misread bases Sequencing artifacts Misaligned reads



At risk: ~1000 germline / Mb (known) 10-20 rare germline / Mb (novel) Source: Low coverage in normal

# **Cross-Center Comparison**

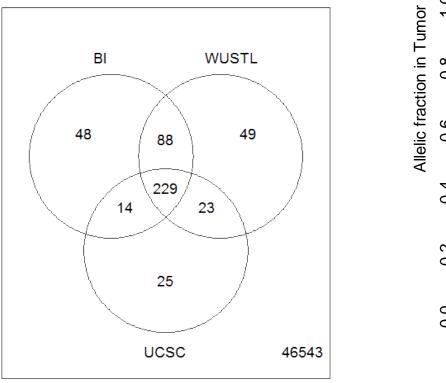
- The project initiated with the goal of comparing, evaluating, and improving mutation calling algorithms
  - Select a set of reference samples
  - Call mutations using different algorithms & compare
- Comparison alone allows only to contrast the callers against each other
  - If caller A makes a call and caller B does not, it is helpful to characterize the difference
  - Is there a difference in heuristics involved?
  - Is there a difference is some statistics of such caller-specific SNVs
  - Ultimately, one needs the ground truth (validation data)

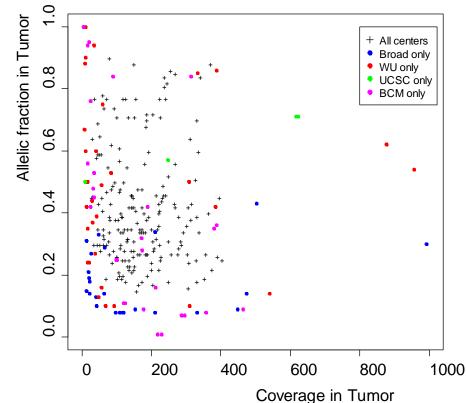
## Data

- For this round of the analysis, the subset of data from Phase III of the project was used
  - 20 Lung Squamous TCGA samples sequenced at Broad (whole-exome)
  - Same sequencing data (distributed between centers as aligned bam files) were called at 4 centers using different algorithms
    - Broad
    - Washington University, Saint Louis
    - UCSC
    - Baylor College of Medicine
  - Resulting callsets shared between the centers for comparison
- In addition, for this work we use RNA-Seq data as a validation dataset
  - Sequenced at UNC for TCGA

### Simple Characterization of Mutation Callers

- Look at shared vs center-specific events
  - There is a large overlap, but there are still many calls made by each center alone
  - The center-specific calls have, in general, different properties
  - Are these specific false-positive modes of each caller or specific strength?

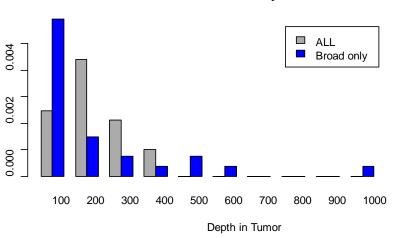




TCGA-33-4532

# Calls vs Coverage

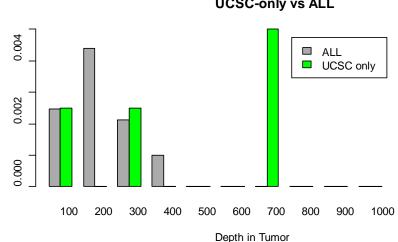
Tendency to call center-specific events at coverages different from • where shared events are located

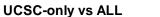


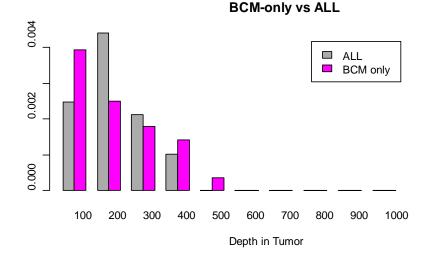
#### **Broad-only vs ALL**

0.006 ALL WUSTL only 0.004 0.002 000 o. 100 200 300 400 500 600 700 800 900 1000 Depth in Tumor

WUSTL-only vs ALL

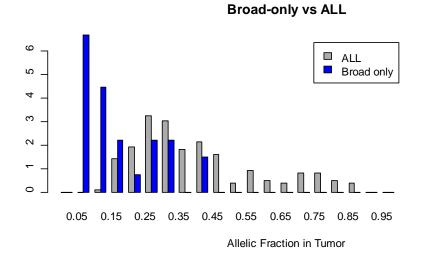


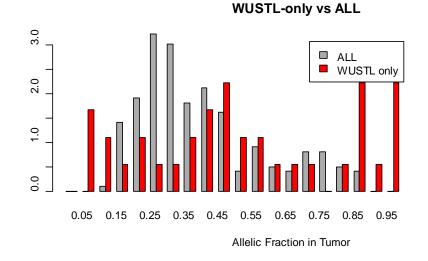




# **Calls vs Allelic Fraction**

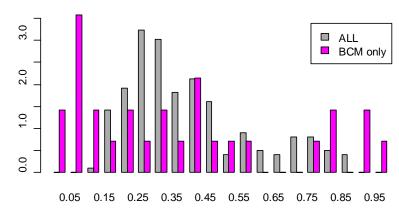
Allelic fraction distribution of center-specific calls differs from that of shared calls





10 ALL ω UCSC only ø 4 2 0 0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95





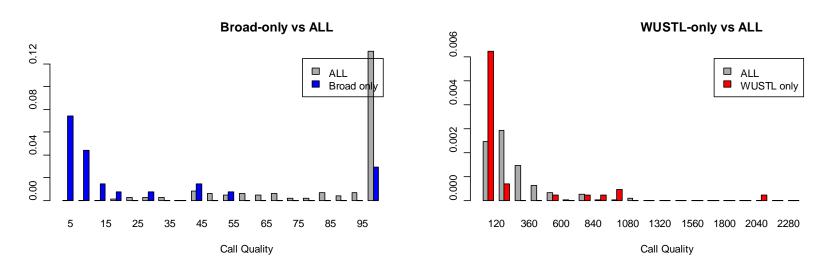
Allelic Fraction in Tumor

**BCM-only vs ALL** 

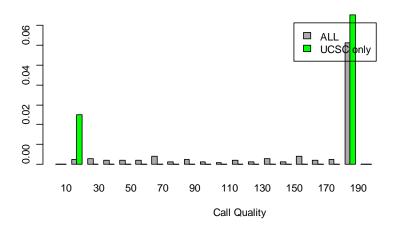
Allelic Fraction in Tumor

# Calls vs Call Quality

• How do callers qualify their own unique calls – are reported qualities meaningful/reliable?



UCSC-only vs ALL



- Some center-specific calls are questionable upon "manual review" (examples follow)
- Many, however, are convincing

# Center-specific call, questionable

- Broad-only, single event at coverage ~1000
  - Questionable alignments in the region; no support in RNA-Seq (all RNA-Seq reads are 0 mapping quality)

| <u> ≝</u> IGV  |                               |  |
|--|-------------------------------|--|
| File View Tracks Help<br>Human hg19                    | ~                             | dr22 v dr22:23,101,340-23,101,531 co 音 < ▶ 🛷 🔲 🗴 🗭 🗸 →   |
|  |                               |  |
|  |                               | p13 p12 p11.2 p11.1 q11.1 q11.2 q11.22 q12.3 q12.1 q12.2 q12.3 q13.1 q13.2 q13.31 q13.32   |
|  |                               |  |
|  | NAME<br>DATA TYP<br>DATA FILI | Image: style="text-align: center;">193 bp           340 bp         23,101,360 bp         23,101,380 bp         23,101,400 bp         23,101,460 bp         23,101,480 bp         23,101,500 bp <t< th=""></t<> |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
| d.flagged.bam Coverage                                 |                               |  |
| LUSC-33-4532.ma.GATKRecalit<br>d.flagged.bam Junctions |                               |  |
|  |                               |  |
|  |                               |  |
|  |                               |  |
| Sequence 🗕   |                               |  |
| RefSeq Genes   |                               |  |
| chr22:23,101,372                                       |                               | 302M of 44\$M  |

# Center-Specific Call, questionable

WUSTL at coverage 5, allelic fraction 0.67
 Likely, a germline event

| 🕌 IGV  |                                |  | J X      |  |  |  |  |  |  |  |  |
|--|--------------------------------|--|----------|--|--|--|--|--|--|--|--|
| Ele View Tradys Help                                   |                                |  |          |  |  |  |  |  |  |  |  |
| Human hg19   | ~                              | drr7 v drr?:97,736,503.97,736,651 Go 🖆 🖌 🕨 🖗 🖪 💥 🖓   | +        |  |  |  |  |  |  |  |  |
|  |                                | p22.2 p21.3 p21.1 p15.2 p14.3 p14.1 p13 p12.2 p11.2 q11.1 q11.22 q21.11 q21.13 q21.3 q22.1 q22.3 q31.1 q31.2 q31.32 q32.1 q33 q34 q35 q36.1 q3 | <u>.</u> |  |  |  |  |  |  |  |  |
|  | NAME<br>DATA TYPE<br>DATA FILE | 150 bp   | →<br>    |  |  |  |  |  |  |  |  |
| LUSC-33-4532-Normal.bam Cove                           |                                | p. 20]   | ^        |  |  |  |  |  |  |  |  |
| LUSC-33-4532-Normal.bam Junc<br>s                      |                                |  |          |  |  |  |  |  |  |  |  |
|  |                                |  | =        |  |  |  |  |  |  |  |  |
| LUSC-33-4532-Normal.bam                                |                                |  |          |  |  |  |  |  |  |  |  |
| LUSC-33-4532-Tumor.bam                                 |                                |  |          |  |  |  |  |  |  |  |  |
| LUSC-33-4532.ma.GATKRecalit<br>d.flagged.bam Junctions |                                |  |          |  |  |  |  |  |  |  |  |
| LUSC-33-4532.ma.GATKRecalib<br>d.flagged.bam           |                                |  |          |  |  |  |  |  |  |  |  |
| Sequence 🔿   |                                |  | 6 C 🔺    |  |  |  |  |  |  |  |  |
| RefSeq Genes<br>chr7:97,736,585                        |                                |  | <br>;=   |  |  |  |  |  |  |  |  |
| 0117197,730,585  |                                | 210M pr 445M   |          |  |  |  |  |  |  |  |  |

# Center-Specific Call, questionable

- BCM in TCGA-66-2777
  - Clearly a germline event

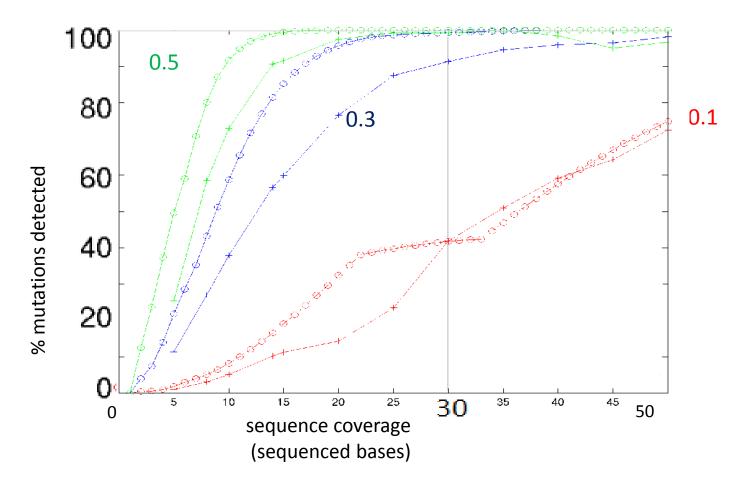
|                                    |                                | chr17                      |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | ^   |
|------------------------------------|--------------------------------|----------------------------|----------|-----------|---------------|-------------------|---------------|----------|----------|--------|---------------|---------|---------|--------|--------------|-------------|---|
|                                    |                                | p13.3 p13.2                | p13.1    | p12       | p11.2         | p11.1             | q11.2         | q12      | q21.1    | q21.31 | q21.32 q21.33 | q22     | q23.1   | q23.3  | q24.2 q24.3  | q25.1 q25.2 | q25.3   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    | FILE                           | - <b>4</b><br>7,579,300 bp |          |           |               | 7 579 400 bp      |               |          | 376      |        | 7 579 500 bp  |         |         |        | 7,579,600 bp |             | *   |
|                                    | NAME<br>DATA TYPE<br>DATA FILE | 1.573,500 BP               |          | 1         |               | 7,579,400 bp<br>  |               |          |          |        | 7,579,500 bp  |         | 1       |        | 7,578,600 Bp |             | · · · · · ·   |
| LUSC-66-2777-Normal.bam Cove       |                                | [0 - 694]                  |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | <u>^</u>  |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
| LUSC-66-2777-Normal.bam Junc       |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
| s                                  |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | •   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         | 1.1    |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | <b>~</b>  |
| LUSC-66-2777-Tumor.bam Cove        |                                | (0 - 630)                  |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | <u>^</u>  |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
| LUSC-66-2777-Tumor.bam Junc        |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            | P.       |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   | <u>п</u>      |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
| LUSC-66-2777.ma.GATKRecalib        |                                | p. 361]                    |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | ×   |
| d.flagged.bam Coverage             |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
| LUSC-66-2777.ma.GATKRecalib        |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
| d.flagged.bam Junctions            |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               | -       |         |        |              |             |   |
|                                    |                                |                            |          |           | _             |                   |               |          |          |        |               |         | 1.1     |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         | -       |        |              |             |   |
| Sequence 🔿                         |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             | <u>~</u>  |
| Sequence $ ightarrow$ RefSeq Genes |                                |                            | CTVSKATG | SHLFGLRFG | 5 Y S G Q Y T | ' K Q S P V S S S | S L P W S P A | РАРААРТР | AAPAPAVE | РААЕРМ | RPAEDPGPDE    | TFWQEID | DPSLMLC | DMAQSF | L P S        | ·           | $\rightarrow \rightarrow \rightarrow \rightarrow \rightarrow$ |
|                                    |                                |                            |          |           |               |                   |               |          |          |        |               |         |         |        |              |             |   |

# WE NEED A LOT OF VALIDATION DATA TO COMPARE THE TOOLS

### Using RNA-Seq as Validation Set

- Independent library construction
- Different protocol
- Same sequencing technology
- It is possible to call mutations (de-novo) from aligned RNA-Seq data
  - Likely a too conservative approach
- Assume that de-novo DNA-Seq mutation calling is sufficiently conservative
  - Weaker evidence from RNA-Seq (than what would be required for a stand-alone de-novo call) can be considered as validation

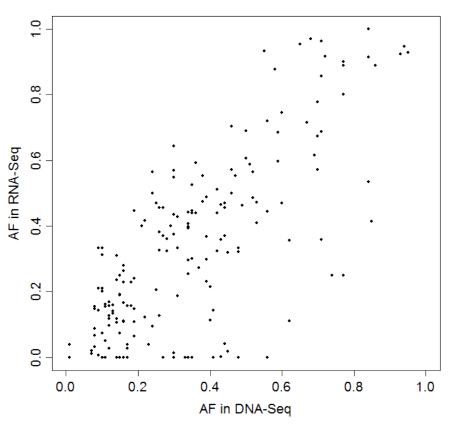




Kristian Cibulskis

## Is Allelic Fraction an Issue?

- Original calls have a range of allelic fractions
- Is it safe to ask for fixed (low) number of observations in RNA-Seq
  - In general, NO
  - However: AF in RNA-Seq and DNA-Seq strongly correlate



LUSC-33-4532, 0.58, 0.75231357069512

# Looking for SNV in RNA-Seq

- Consider every called mutation site with coverage in RNA-Seq above N as "covered"
- If covered site has at least two reads with alt. allele in RNA-Seq, consider it "validated"

| RNA-Seq T cov. >=5  | validated.pct.covered<br>81.7 |                  | covered<br>186 | center<br>BI |
|---------------------|-------------------------------|------------------|----------------|--------------|
| RNA-Seq T cov. >=10 | validated.pct.covered<br>87.3 | validated<br>131 | covered<br>150 | center<br>BI |
| RNA-Seq T cov. >=20 | validated.pct.covered<br>93.6 | validated<br>102 | covered<br>109 | center<br>BI |

# Conclusions

- A framework is established within TCGA for evaluating and improving mutation calling algorithms
- We are working on validating mutations:
  - Using additional experiments in the sequencing centers (but this may be only partial validation)
  - based on RNA-seq after correcting for the power to detect the mutation

# Acknowledgments

#### Broad

Gad Getz Kristian Cibulskis Rui Jing Alex Ramos Carrie Sougnez Peter Hammerman Scott Carter

### WUSTL

Li Ding Mike McLellan Ken Chen Xian Fan

### UCSC

David Haussler Chris Wilks Singer Ma Zack Sanborn Rachel Harte Daniel Zerbino Jing Zhu

#### **Baylor**

David Wheeler Jennifer Drummond Kyle Chang

#### UNC

Neil Hayes Matthew Wilkerson

### TCGA research Network