



# Identification of Low Prevalence Somatic Mutations in Heterogeneous Tumor Samples



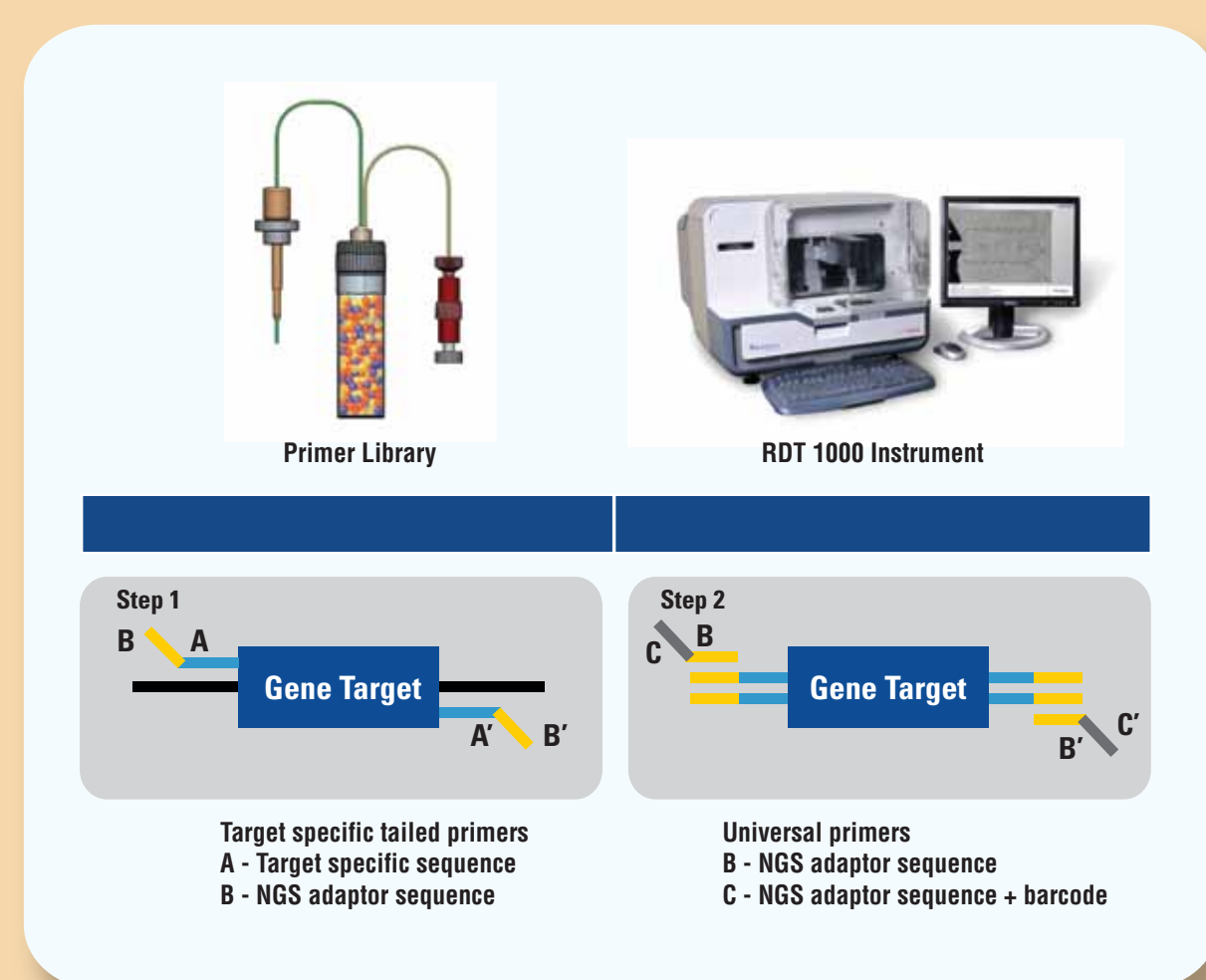
Olivier Harismendy<sup>1,2</sup>, Lei Bao<sup>1</sup>, Steve Kotsopoulos<sup>3</sup>, Sophie Rozenhak<sup>1,2</sup>, Jeff Olson<sup>3</sup>, Masakazu Nakano<sup>1,2</sup>, Brian Crain<sup>1</sup>, Stephanie Pond<sup>4</sup>, Karen Messer<sup>1</sup>, Richard Schwab<sup>1</sup>, Mark S. Chee<sup>4</sup>, Darren R. Link<sup>3</sup>, Kelly A. Frazer<sup>1,2,5</sup>.

**ABSTRACT:** High throughput sequencing enables the digital measurement of each allele in a DNA sample. This provides an ideal method to interrogate mutations present in heterogeneous samples such as solid tumors in which clonal selection or contamination with stroma can hinder the identification of important somatic mutations. We developed an ultra-deep targeted sequencing (UDT-Seq) assay to screen 42 cancer genes via microdroplet-based PCR (RainDance Technologies) and direct sequencing of the amplicons on the Illumina GA. This UDT-Seq library interrogates ~86 kb of DNA located in cancer mutational hotspots (87% of all COSMIC database entries) and ~23 kb located in exons sequenced in HapMap samples for the assay calibration and performance evaluation. We devised a statistical filtering of the mutations by using both experimental estimation of the sequencing error rate and training with a calibration sample. We measured the performance of our assay by processing 4 blends of 4 HapMap samples, interrogating 158 SNPs with known prevalence in each blend. The sensitivity and specificity of our method is >88% and >99% respectively for mutations present at 1% or greater. We next interrogated 4 cancer samples (xenografts, 2 of which with matching primary samples) from 4 different fresh frozen tissues. We were able to detect and validate low-prevalence somatic mutations in all samples of which some are well-known driver mutations. Finally, we analyzed the robustness of the detection and prevalence measurement after performing whole genome amplification and show that WGA leads to an underestimation of the mutant allele for mutations present at 5% prevalence or lower. Featuring a streamlined sample preparation to interrogate a large number of bases, this assay is well suited for clinical applications to study clonal selection in cancer progression or treatment with sub-optimal heterogeneous cancer samples.

## UDT-Seq Assay and Experimental Design

### Assay Design

- Microdroplet-based PCR
- 2-Step Tailed Primer Assay
- No sequence library preparation for streamlined implementation in clinic
- Direct sequencing 2\*125 bp to control error rate



**The UDT-Seq laboratory workflow.** The DNA samples and the chimeric targeting primer are merged on the RDT1000 instrument. Following the PCR, the emulsion is broken and after purification the resulting samples are re-amplified using universal primers which contain the sequence required for Illumina GA sequencing.

### Analysis Strategy

1. Estimate the sequencing error rate
2. Identify candidate mutations in both training and tested samples
3. Determine the significance cutoff using known SNPs in the training sample
4. Apply this cutoff to the candidate mutations in the tested sample

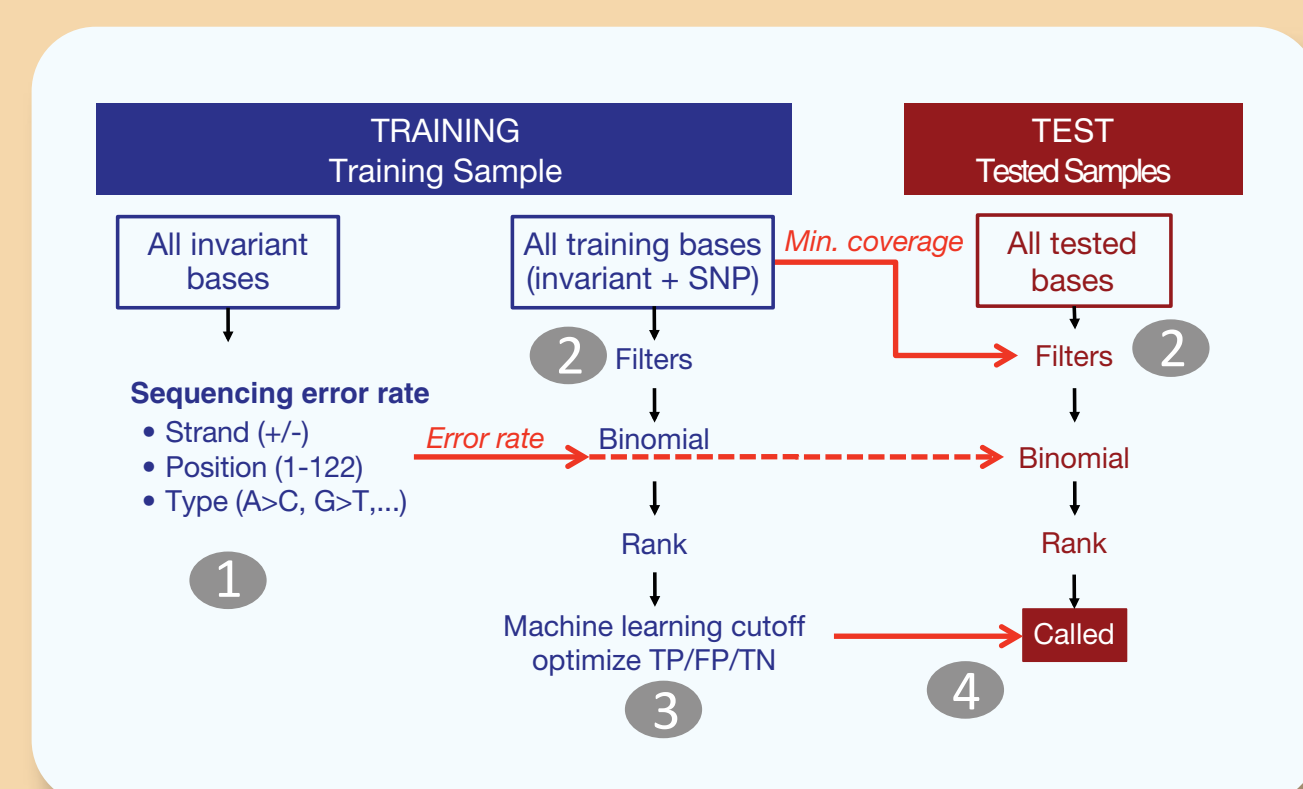
### Targets Selection

#### Cancer Hotspots (86.7 kb)

- COSMIC v44 : 9,935
- 2% observed more than 100 times
- Known coordinates
- Selected 5271 mutations in 42 genes
- Substitutions or short indel
- 87% of all valid COSMIC entries
- 96% observed less than 5 times
- 518 amplicons (200 bp)

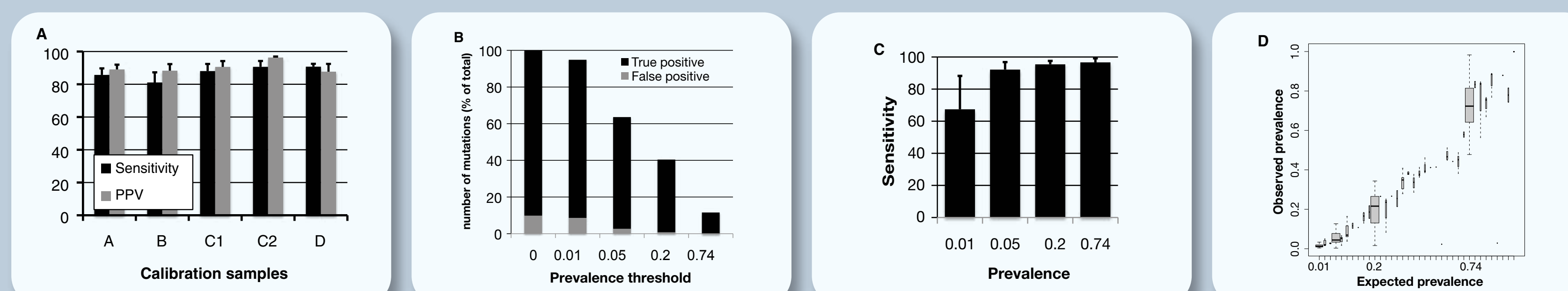
#### Calibration SNPs (23.2 kb)

- Calibration samples are a blend of 4 known samples at 1%, 5%, 20%, 74%
- 158 calibration SNPs are known SNPs, homozygous alternate in 1 out of 4 samples
- Calibration targets are 200 bp amplicons around SNPs



**Flow Diagram of the analysis strategy:** The four steps are indicated in grey: Values generated in the training sample and applied to the validation are connected in red

## Performance Evaluation



- At 1% detection:**
- 99.99% specificity
  - 88% sensitivity
  - 90% positive predictive value

73% of false positives are <5% prevalence

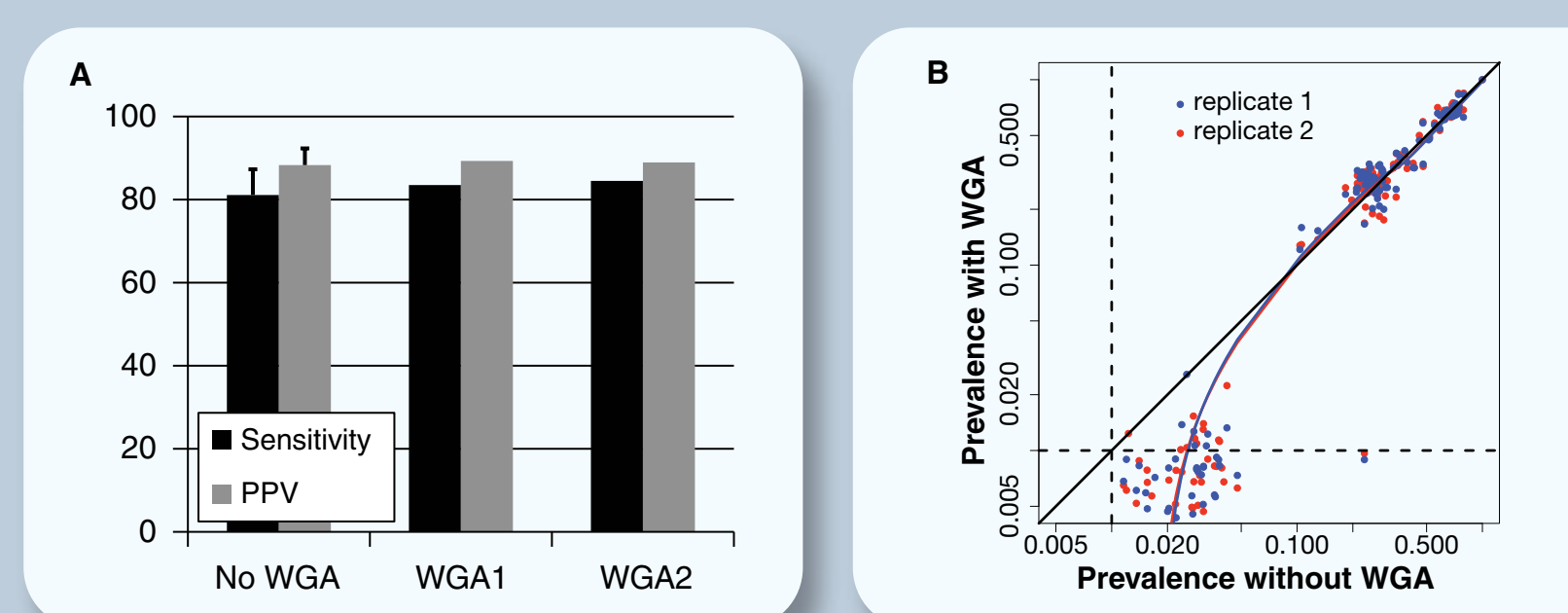
Sensitivity increases with prevalence

Expected ~ Observed (r=0.96)

(A) The sensitivity and positive predictive value (PPV) of UDT-Seq for each of the 5 calibration datasets. The error bars represent the standard deviation of the values obtained from different training schemes. (B) Proportions of True Positive (black) and False Positive (grey) mutations at different prevalence thresholds (x-axis) across all tested calibration samples from different training

schemes. (C) Average sensitivity estimation for the different categories of calibration SNPs. The error bars represent the standard deviation over all tested calibration samples analyzed with all training schemes. (D) Expected prevalence of the calibration SNPs (x-axis) are highly correlated (r=0.96) with the observed prevalence (y-axis) across all calibration samples

## Effect of Whole Genome Amplification



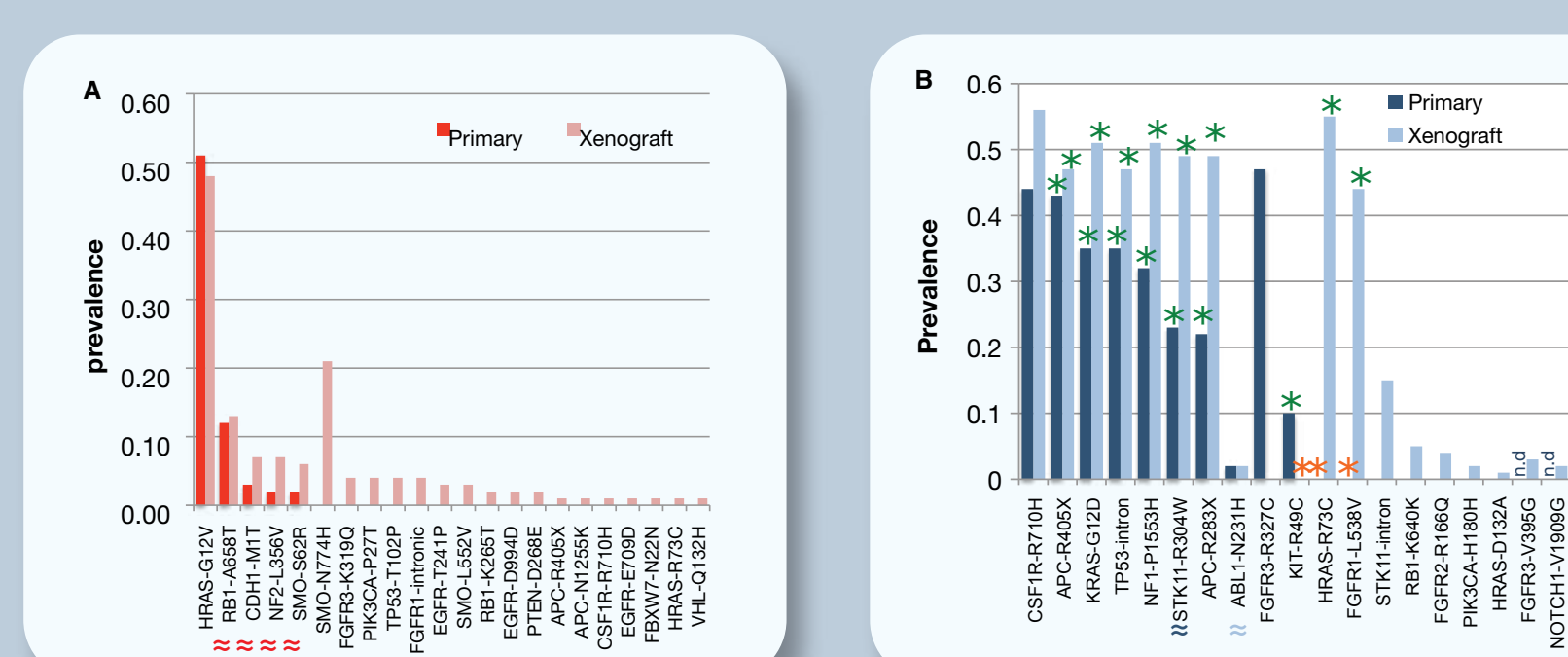
Although still significantly called, low prevalence (<5%) mutations are underestimated in WGA sample

### Performance of UDT-Seq after Whole Genome Amplification.

(A) The Sensitivity and PPV were calculated on calibration sample CAL-B without amplification (error bars correspond to the standard deviation after training with sample CAL-C and CAL-D) or with whole genome amplification, in duplicate. (B) The prevalence of the calibration SNPs identified in CAL-B without amplification (x-axis) is plotted against the prevalence estimated from the WGA amplified sample replicates (red and blue).

## Application to Cancer Samples

### Comparison Between Primary and Xenograft Samples



8 mutations identified in both primary and xenograft samples

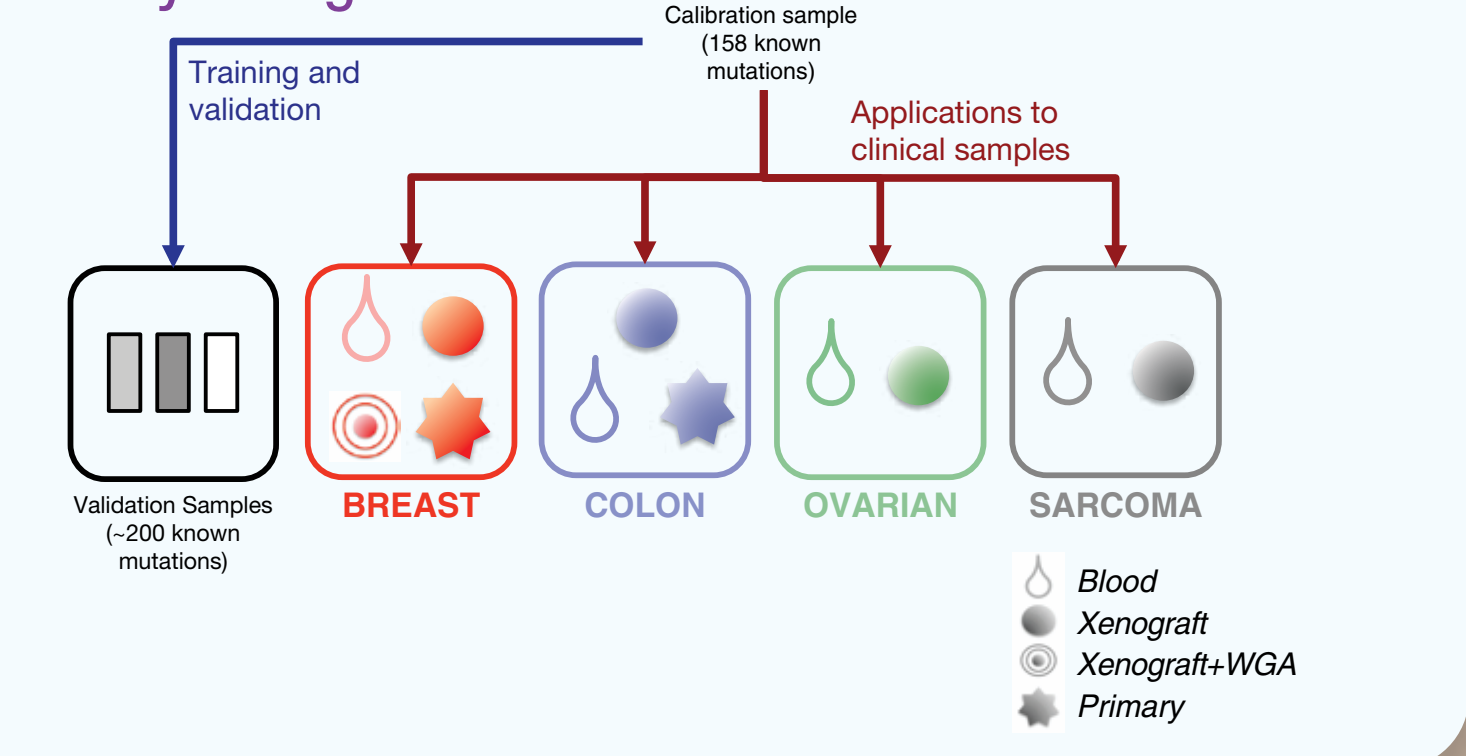
- Well known colon cancer driver (KRAS-G12D, APC-R405X)
- Cell population changes during xenograft process:
  - prevalence changes 25%->50% (APC and STK11)
  - 2 mutations are lost during xenograft
  - >3 mutations are gained in the xenograft

- HRAS-G12V prominent in both xenograft and primary
- Additional xenograft specific mutations

\* SNaPshot confirmed  
 \* SNaPshot not significant  
 \* SNaPshot not confirmed  
 ≈ manual estimation (following color legend)

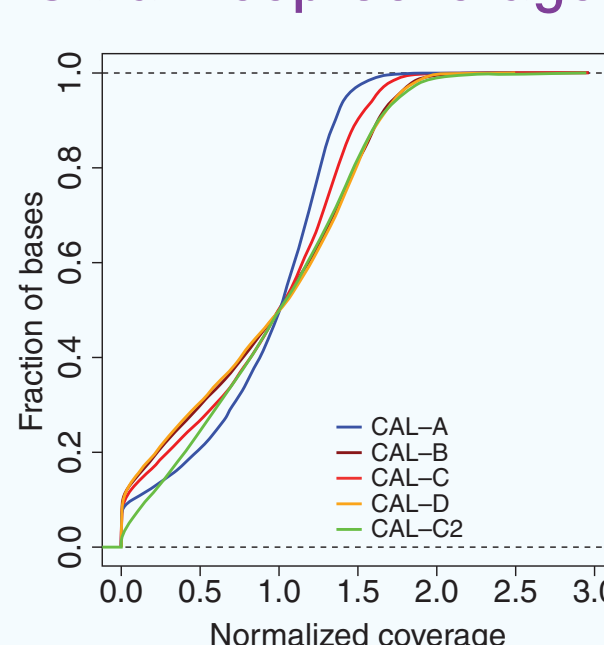
**Application of UDT-Seq to Cancer Samples.** (A) Histogram of the prevalence in all the mutations identified in the breast primary (red) or xenograft (pink) cancer sample. (B) Histogram of the prevalence of all the mutations identified in the colon primary (blue) or xenograft (light blue) cancer sample. For both A and B panels, SNaPshot validation results are indicated by the color-coded asterisk (see legend), manual evaluation is indicated by (≈) and n.d. means not covered.

### Study Design



(1) Use of known calibration sample to validate the assay and to assess performance at known SNPs and (2) Apply the assay to four clinical samples: 1 breast primary metastatic carcinoma, 1 colon enteric adenocarcinoma, 1 ovarian serous adenocarcinoma and 1 small intestine sarcoma

### Ultra-Deep Coverage



~24,000x average coverage depth 79.6% of bases within 0.5-2 fold of the mean

1. Moores UCSD Cancer Center and 2. Department of Pediatrics and Rady Children's Hospital, University of California San Diego, 9500 Gilman Drive, La Jolla CA 92093. 3. RainDance Technologies, 44 Hartwell Avenue, Lexington MA 02421. 4. Prognosis Biosciences, 505 Coast Blvd, La Jolla CA 92037. 5. Institute for Genomic Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla CA 92093.

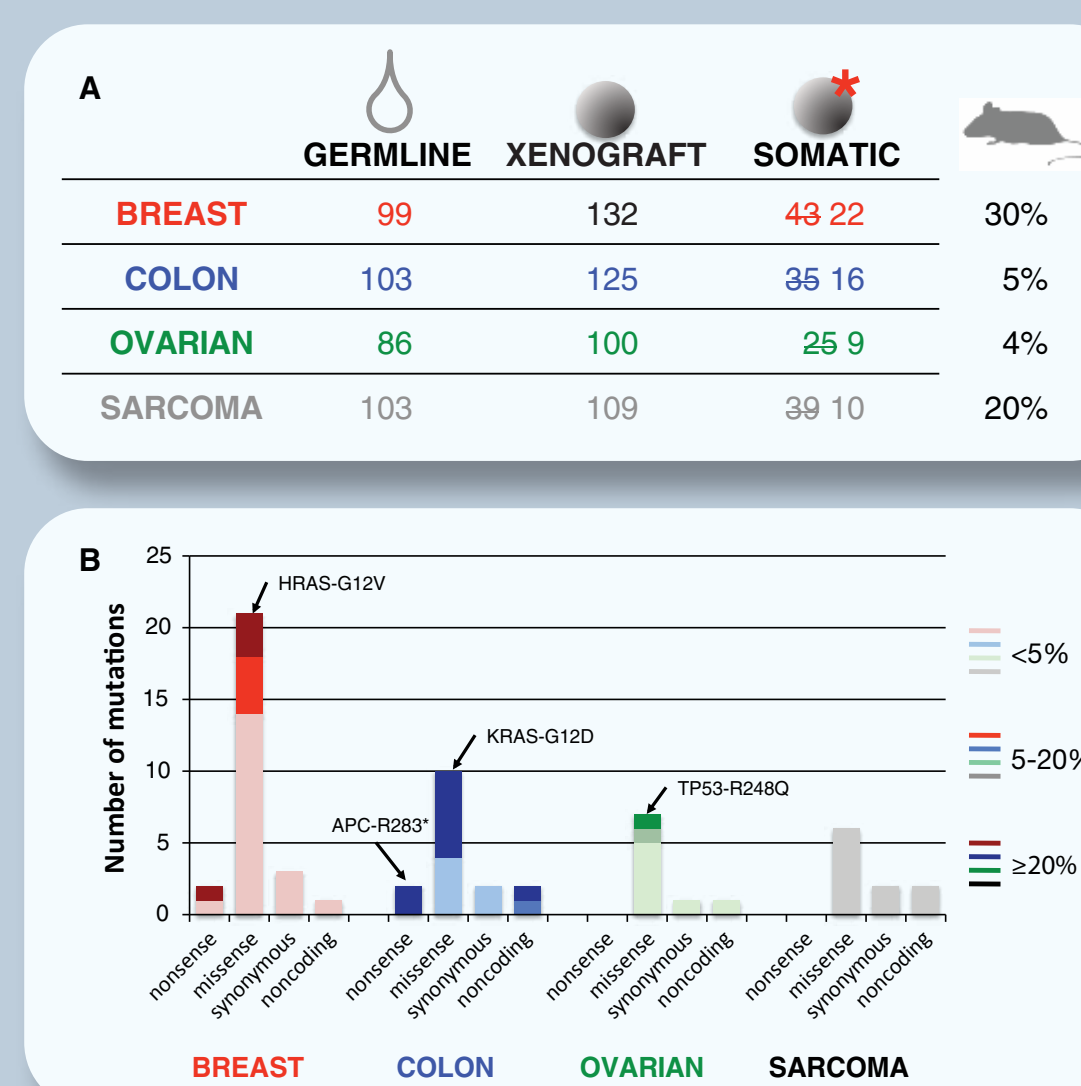
## Perspectives

- Development**
  - Formalin Fixed Paraffin Embedded Samples Ion Torrent/MiSeq for faster turn around
- Patient management**
  - Monitoring: establish mutation signature, monitor by digital PCR
  - Guide Therapy: molecular characterization of the tumor
- Clinical research**
  - Support targeted trial: stratify by mutation
  - Retrospective trial: marker of prognosis
- Basic research**
  - Identify new driver mutations in known cancer genes
  - Study clonal selection (in vivo/in vitro)
  - Study mechanisms of resistance

## Conclusions

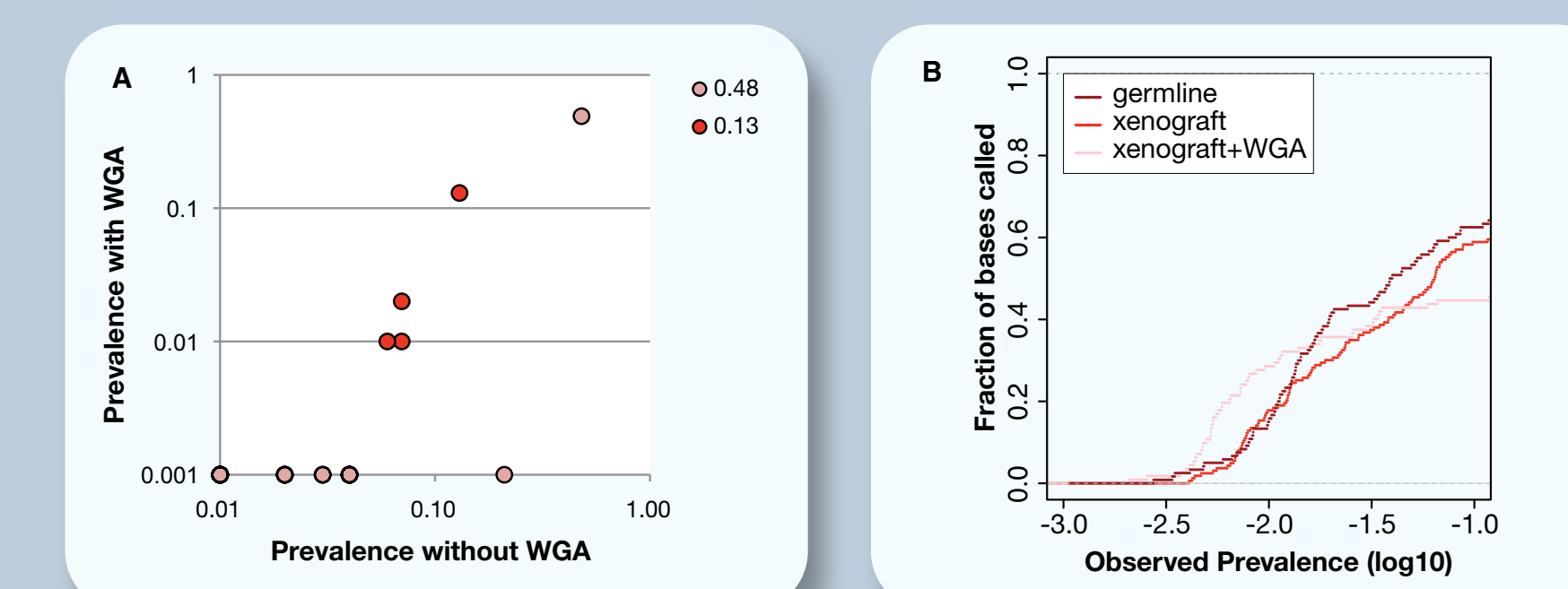
- Enables the discovery of new mutations
- Sensitive down to 1% prevalent mutations
- Comprehensive screening more than 5,000 known cancer mutations, 87,000 base pairs total
- Robust due to simultaneous analysis of a calibration sample
- Streamlined for clinical implementation

## Analysis of Xenografts Samples



- 75/132 somatic mutations are mouse contaminant
  - Majority of low prevalence mutations in the xenografts
- (A) Table indicating the number of mutations identified by UDT-Seq in the germline and xenograft samples, the number of resulting somatic mutations before (crossed) and after removal of mouse-human mismatches. (B) Distribution of the mutations and their functional consequences in the four xenografts samples at low (light colors) intermediate (medium colors) and high (dark colors) prevalence.

## Effect of Whole Genome Amplification



Whole Genome Amplification induces bias and under-estimation of the low prevalence (<5%) mutations

(A) Comparison of prevalence of 22 mutations identified without WGA amplification (x-axis) and with WGA amplification (y-axis) of a breast xenograft sample. Mutations identified manually in the WGA sample are indicated in red. (B) Cumulative distribution of the candidate mutations in the germline, xenograft with WGA and xenograft WGA of the breast cancer patient shows an excess of rare prevalence due to WGA bias.

UDT-Seq can detect somatic mutations, measure their prevalence and enable the study of clonal selection