Title

Assessing aneuploidy with repetitive element sequencing.

Authors: Christopher Douville,[1,2,3,4] Joshua D. Cohen,[1,2,3,4,5] Janine Ptak,[1,2,3,4] Maria Popoli,[1,2,3,4] Joy Schaefer,[1,2,3,4] Natalie Silliman,[1,2,3,4] Lisa Dobbyn,[1,2,3,4] Robert E, Schoen,[6,7] Jeanne Tie,[8,9,10,11] Peter Gibbs,[8,9,10] Michael Goggins,[12] Christopher L. Wolfgang,[13] Tian-Li Wang,[2,3,12,14] Ie-Ming Shih,[12] Rachel Karchin,[5,13,14] Anne Marie Lennon,[2,3,13,14,16] Ralph H. Hruban,[12,14] Cristian Tomasetti,[1,2,3], Chetan Bettegowda,[1,2,3,17] Kenneth W. Kinzler,[1,2,3,*] Nickolas Papadopoulos,[1,2,3] Bert Vogelstein,[1,2,3,4,*]

Affiliations:

[1]Ludwig Center for Cancer Genetics and Therapeutics, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

2Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[3]Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[4] Howard Hughes Medical Institute, Baltimore, MD 21287, USA.

[5]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA.

[6]Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15260, USA.

[7]Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA 15260, USA.

[8]Division of Personalized Oncology, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia. 1

[9]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, VIC 3010, Australia.

[10]Department of Medical Oncology, Western Health, Melbourne, VIC 3021, Australia.

[11]Department of Medical Oncology, Peter MacCallum Cancer Center, Melbourne, VIC 3000, Australia.

[12]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA.

[13]Department of Surgery, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA.

[14] Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[15]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA.

[16]Department of Medicine, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA.

[17]Department of Neurosurgery, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA.

Abstract

We report a sensitive PCR -based assay that can detect aneuploidy in samples containing as little as 3 picograms of DNA. Using a single primer pair, we amplified ~750,000 amplicons distributed throughout the genome Aneuploidy was detected in 49% of liquid biopsies from a total of 883 non-metastatic cancers of eight different types. Combining aneuploidy with somatic mutation detection and eight standard protein biomarkers yielded a median sensitivity of 80% at 99% specificity.

As a result of drastic reductions in costs, whole genome sequencing (WGS) is now commonly used to detect chromosome copy number variations, also known as aneuploidy [1]. Identifying the presence of aneuploidy has a broad range of diagnostic applications including non-invasive prenatal testing (NIPT) [2], preimplantation genetic diagnosis [3], evaluation of congenital abnormalities [4], and cancer diagnostics [5].

Shallow (0.1x-1x) WGS is employed for aneuploidy detection in a large number of commercially available tests [6]. WGS is typically employed in NIPT, where a relatively high fraction (5-25%) of the total DNA is derived from the fetus [7]. A companion diagnostic is frequently used to estimate the fetal fraction and NIPT is often not performed when the fraction of fetal DNA is less than 4% [8, 9]. Sequencing depth becomes a major issue for the assessment of aneuploidy in cell-free DNA from patients with cancer, where the fraction of DNA derived from cancer cells is often much less than 1% of the total input DNA [10].

Amplicon-based methods using sequence-specific primers have been proposed as an alternative to WGS for the assessment of aneuploidy [11-13]. Amplicon-based protocols offer many advantages over WGS (or exome sequencing), including a simpler workflow that does not require library construction, a reduced requirement for input DNA, and a simplified computational analysis. Here, we report a substantially improved amplicon-based approach to detect the presence of aneuploidy, named the Repetitive Element AneupLoidy Sequencing System (REAL-SeqS). Using a single PCR primer pair, REAL-SeqS amplifies ~750,000 genomic loci with an average size of 88 base pairs spread throughout the genome.

REAL-SeqS introduces six key innovations:

- higher sensitivity with less sequencing than previously reported technologies
- increased spatial coverage throughout the genome, enabling the detection of microdeletions and microamplifications.
- an improved machine-learning-based algorithm for interpreting the data
- reduced requirement for input DNA

- concomitant detection of contaminating cellular DNA in samples of cell-free DNA
- improved detection of aneuploidy in cell-free DNA.

The FAST-SeqS approach described in Kinde et al was the first aneuploidy detection method to use a single primer pair to amplify numerous repetitive long interspersed nucleotide elements (LINEs) spread throughout the genome [11]. However, due to the low genomic density of these amplicons (a total of 38,000 across the entire genome), its power to detect focal amplifications and deletions (<5MB) was limited. Additionally, FAST-SeqS amplicons ranged in size from 120-145 bps, which was sub-optimal for assessing cell-free DNA, which has an average size of ~140 bp. Accordingly, FAST-SeqS was only able to detect aneuploidy in 22% of liquid biopsy samples containing more than 1% of tumor-derived DNA [14].

Based on the limitations described above, we attempted to identify a single primer pair that could amplify far more than 38,000 amplicons of a size far less than 120 to 145 bp. To generate a list of candidate primers, we first calculated the frequency of all possible 6-mers (4^6 = 4096) within the RepeatMasker track of hg19. Next, we calculated the frequency of all possible 4-mers (4^4 = 256) within 75 bp upstream or downstream from the 6-mers. Joining the 6-mers with the 4-mers generated 2,097,152 candidate pairs. We narrowed these pairs based on the number of unique genomic loci expected from their PCR-mediated amplification, the average size between the 6-mer and its corresponding 4-mers, and the distribution of these sizes, aiming for a unimodal distribution. This filtering criteria generated 7 potential k-mer pairs, leading to the design of 7 primer pairs that incorporated these k-mer pairs at their 3-ends. Two of these primer pairs (REAL1 and REAL2) outperformed the remaining 5 primers when assessed experimentally by the number of unique loci that were amplified and the size distribution of the amplicons. After further experimental testing of REAL1 and REAL2 on 100 normal samples, the REAL1 primer pair was chosen for the experiments reported herein. The average amplicon size of REAL1 was 88 base pairs (Supplementary Figure 1). Details of the primer selection methods, experimental procedures, new analytic techniques, and work flow diagram are described in Supplementary Text and Supplementary Figure 2.

In the most common form of NIPT, detection of a gain or loss of a chromosome (e.g., chromosome 21 in Down Syndrome) is the goal. We used WGS (Supplementary Table 2), FAST-SeqS (Supplementary Table 3), and REAL-SeqS (Supplementary Table 4) to assess performance on a collection of synthetic samples for DNA admixtures typically encountered in NIPT, i.e., when the fraction of fetal DNA was 5% (pseudocode used to generate synthetic samples Supplementary Figures 3 and 4). To ensure that these comparisons were intrinsic to the sequencing data rather than to the computational algorithm used to analyze the data, we calculated performance using simple z score comparisons (Supplementary Text). We reported results in total reads needed for all three approaches assuming single-end 100 bp reads and accounting for differences in alignment rates and filtering criteria typically used (Supplementary Tables 2, 3, and 4), REAL-SeqS consistently achieved higher sensitivity at lower amounts of sequencing. For example, REAL-SeqS had 98.5% sensitivity (at 99% specificity) for monosomies and trisomies at a 5% cell fraction, while WGS and FAST-SeqS had 93.9% and 81.1% sensitivity respectively (Figure 1A).

Another important aspect of assays for copy number variation is the detection of relatively small regions which are deleted or amplified. For example, the DiGeorge Syndrome deletions are often as small as 1.5 Mb [15]. For a 5% deletion-containing cell fraction, REAL-SeqS had 75.0% sensitivity for the 1.5 Mb DiGeorge deletion (at 99% Specificity) while WGS and FAST-SeqS had 19.0% and 29.0% sensitivity, respectively (Figure 1B).

The detection of amplifications, such as those on *ERBB2* in breast cancer, are critical for deciding whether patients should be treated with trastuzumab or other targeted therapies. Following the same protocol as above, we generated synthetic samples with focal amplifications of the ~42 Kb ERBB2 gene (20 copies) for WGS, FAST-SeqS, and REAL-SeqS. REAL-SeqS could detect such amplifications in the synthetic samples with significantly less sequencing than could WGS or Fast-SeqS. For a 1% cell fraction, REAL-SeqS had a 91.0% sensitivity while WGS had 50.0% (Figure 1C). FAST-SeqS did not have enough spatial coverage in this genomic region to detect *ERBB2* amplifications.

Reliably detecting aneuploidy in only a few pg of DNA is necessary for preimplantation diagnostics as well as forensic applications. In preimplantation diagnosis, a few cells picked from a blastocyst are used to assess copy number variations, such as those responsible for Down Syndrome. To test the limit of detection of REAL-SeqS with respect to input DNA, we analyzed trisomy 21 samples at input DNA concentrations ranging from 3-34 pg (Supplementary Figure 5 and Supplementary Data 2). Trisomy 21 was detected in all these samples, even those from 3 pg of DNA, representing half of a diploid cell. No chromosome arms other than chromosome 21 were found to be aneuploid in the Trisomy 21 samples. No chromosome arms, including chromosome 21, were found to be aneuploid in the euploid controls used in these experiments. The reduced requirement for input DNA also enables retrospective testing of samples from biobanks for either aneuploidy or identification purposes (using SNPs within the amplified repeated sequences; see Supplementary Material).

Plasma "cell-free" DNA is often contaminated with DNA that has leaked out of leukocytes, either during phlebotomy or preparation of plasma. This contaminating leukocyte DNA can reduce the sensitivity of aneuploid testing for plasma samples because leukocytes are not derived from either fetal cells (in NIPT) or cancer cells (in liquid biopsies). Leukocyte DNA (gDNA) has an average size of >1000 bp while cell-free plasma DNA has an average size of < 160 bp. DNA size impacts PCR efficiency and long amplicons may not be present in cfDNA. REAL-SeqS enables the detection of leukocyte DNA contamination by virtue of the differently-sized amplicons generated with REAL1 primers. We identified 1241 amplicons typically present in gDNA but not cfDNA (Supplementary Data 3). Reads at these amplicons thereby indicates leukocyte contamination in plasma samples (Supplementary Data 3).

Through mixing of leukocyte DNA with cell-free plasma DNA, we were able to demonstrate that samples containing >4% of leukocyte DNA could be detected with REAL-SeqS (Supplementary Table 5).

DNA from cancer cells is shed into the bloodstream, fostering the analysis of cell-free DNA in plasma ("liquid biopsies") to detect the presence of cancers. Several features of cancer DNA, including point mutations, aberrant DNA methylation, and aneuploidy have been used to assess liquid biopsies. Because aneuploidy is a feature of virtually every cancer type (>90%), it is well-suited for this purpose [14, 16].

REAL-SeqS was used to detect aneuploidy in cell-free plasma DNA from 883 patients harboring surgically resectable cancers of 8 different cancer types (ovary, colorectum, esophagus, liver, lung, pancreas, stomach, and breast). Each plasma sample was given a REAL-SeqS score based on a machine-learning-based algorithm described in the Supplementary Text. Aneuploidy was scored in the samples from cancer patients at a threshold of 99% specificity derived from the analysis of 1348 plasma samples from healthy individuals (Supplementary Data 4 and 5). The plasma samples from cancer patients had previously been analyzed for somatic point mutations and small insertions or deletions using a sensitive mutation detection technique based on 61 genomic regions that are frequently altered in cancer [10]. Mutations in the plasma samples were also scored at a threshold of 99% specificity.

Overall, we found that aneuploidy was detected more commonly than mutations (49% and 34% of 883 samples, respectively) in plasma samples from cancer patients ($P < 10^{-20}$, one sided binomial test) With respect to tissue type, aneuploidy was detected more commonly than mutations in samples from patients with cancers of the esophagus, colorectum, pancreas, lung, stomach, and breast, (all P-values <0.01), less commonly in ovary (P=0.048), and equally commonly in liver cancer (Figure. 2A). Importantly, aneuploidy was detected in 242 (42%) of plasma samples in which mutations were not detected, and conversely, mutations were detected in 112 (25%) of samples in which aneuploidy was not detected. Higher amounts of tumor DNA were associated with higher sensitivities. Aneupoidy was detected in 89 of 94 (95%) samples of high ctDNA content (somatic mutation allele frequency >1%). Mutations in the plasma originating from clonal hematopoiesis of indeterminant potential (CHIP), rather than from cancer cells, has confounded previous analyses of mutations in cell-free DNA. This confounder was mitigated with aneuploidy detection; 0 of 17 samples that had CHIP mutations were positive for aneuploidy. We also tested leukocyte DNA from 18 patients whose plasma samples scored positive for aneuploidy with REAL-SeqS; only one of these leukocyte samples was aneuploid as assessed by REAL-SeqS.

Either mutations or aneuploidy were detected in 551 of the 883 plasma samples (62%). This performance is likely an underestimate of what could be achieved: more mutations might have been detected if more amplicons were sequenced and additional aneuploidy might have been identified at a greater sequencing depth. However, in practice there must be a balance between sensitivity and cost, thus limiting the amount of sequencing that can be performed in a screening setting.

Finally, we evaluated the ability of a multi-analyte test, combining protein biomarkers for cancer with aneuploidy and mutations in the cohort described above (detailed methods in Supplementary Text). Eight protein biomarkers were evaluated in these 883 cancer samples as previously described [10]. We scored plasma samples using a logistic regression model maintaining an aggregate specificity of 99% (Supplementary Data 5, 6, and 7). In the plasma samples of patients harboring cancers of seven tissue types (liver, ovary, pancreas, esophagus, stomach, colorectal, and lung), the median sensitivity was 80% (range 77% to 97%), while in breast cancers, it was 38% (Figure 2B).

In summary, REAL-SeqS is exceedingly simple to perform, requires only a single primer pair, and is relatively sensitive and cost effective (~$110 per assay). We anticipate that it will be used to assess aneuploidy in a variety of clinical contexts.

Acknowledgements

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, The Virginia and D.K. Ludwig Fund for Cancer Research, The Commonwealth Fund, Burroughs Wellcome Career Award for Medical Scientists, The Honorable Tina Brozman Foundation, Gray Foundation, Susan Wojcicki and Dennis Troper, the Rolf Foundation, The Templeton Foundation, and NIH grants T32-GM007309, NCI grants U01CA230691-01, P50CA228991, U01CA200469, R37 CA230400-01, CA62924, CA210170 .

Author Contributions

C.D., N.P., K.W.K., and B.V. designed research; C.D., N.P., K.W.K., and B.V. performed research; C.D., J.D.C., J.P., M.P., J.S., N.S., L.D., R.E.S., J.T., P.G., M.G., C.L.W., T.L.W., I.M.S., R.K., A.M.L., R.H.H., C.B., C.T., N.P., K.W.K., and B.V. contributed new reagents/analytic tools; C.D. and B.V. analyzed data; and C.D. and B.V. wrote the paper.

Competing Interests:

Conflicts BV, KWK, & NP are members of the Scientific Advisory Board of Sysmex and are founders of Thrive, Personal Genome Diagnostics, and advise Sysmex. KWK & BV advise Eisai, CAGE Pharma, Neophore, and Morphotek, and BV is also an advisor to Nexus. CB is a consultant for Depuy-Synthes. The companies named above, as well as other companies, have licensed previously described technologies related to the work described in this paper from Johns Hopkins University. CD, RK, CT,and JC, along with BV, KWK, and NP, are inventors on these technologies. Some of these licenses are or will be associated with equity or royalty payments to CD, JC, BV, KWK, and NP. Additional patent applications on the work described in this paper may be filed by Johns Hopkins University. The terms of all these arrangements are being managed by Johns Hopkins University in accordance with its conflict of interest policies.

1. Raman, L., Dheedene, A., De Smet, M., Van Dorpe, J. & Menten, B.r. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic acids research* **47**, 1605-1614.
2. Dheedene, A. et al. Implementation of noninvasive prenatal testing by semiconductor sequencing in a genetic laboratory. *Prenatal diagnosis* **36**, 699-707.
3. Deleye, L. et al. Shallow whole genome sequencing is well suited for the detection of chromosomal aberrations in human blastocysts. *Fertility and sterility* **104**, 1276-1285. e1271.
4. Liang, D. et al. Copy number variation sequencing for comprehensive diagnosis of chromosome disease syndromes. *The Journal of Molecular Diagnostics* **16**, 519-526 (2014).
5. Leary, R.J. et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Science translational medicine* **4**, 162ra154-162ra154 (2012).
6. Roberts, N.J. et al. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer discovery*, CD-15-0402 (2015).
7. Suzumori, N. et al. Fetal cell-free DNA fraction in maternal plasma is affected by fetal trisomy. *Journal of human genetics* **61**, 647 (2016).
8. Gromminger, S. et al. Fetal aneuploidy detection by cell-free DNA sequencing for multiple pregnancies and quality issues with vanishing twins. *Journal of clinical medicine* **3**, 679-692.
9. Nicolaides, K.H., Syngelaki, A., Ashoor, G., Birdir, C. & Touzet, G. Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population. *American journal of obstetrics and gynecology* **207**, 374. e371-374. e376 (2012).
10. Cohen, J.D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926-930.
11. Kinde, I., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. FAST-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing. *PloS one* **7**, e41162.
12. Grasso, C. et al. Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data. *The Journal of Molecular Diagnostics* **17**, 53-63.
13. Tan, C. et al. A multiplex droplet digital PCR assay for non-invasive prenatal testing of fetal aneuploidies. *Analyst* (2019).
14. Douville, C. et al. Detection of aneuploidy in patients with cancer through amplification of long interspersed nucleotide elements (LINEs). *Proceedings of the National Academy of Sciences* **115**, 1871-1876.
15. Packham, E.A. & Brook, J.D. T-box genes in human disorders. *Human molecular genetics* **12**, R37-R44 (2003).
16. Knouse, K.A., Davoli, T., Elledge, S.J. & Amon, A. Aneuploidy in cancer: Seq-ing answers to old questions. (2017).

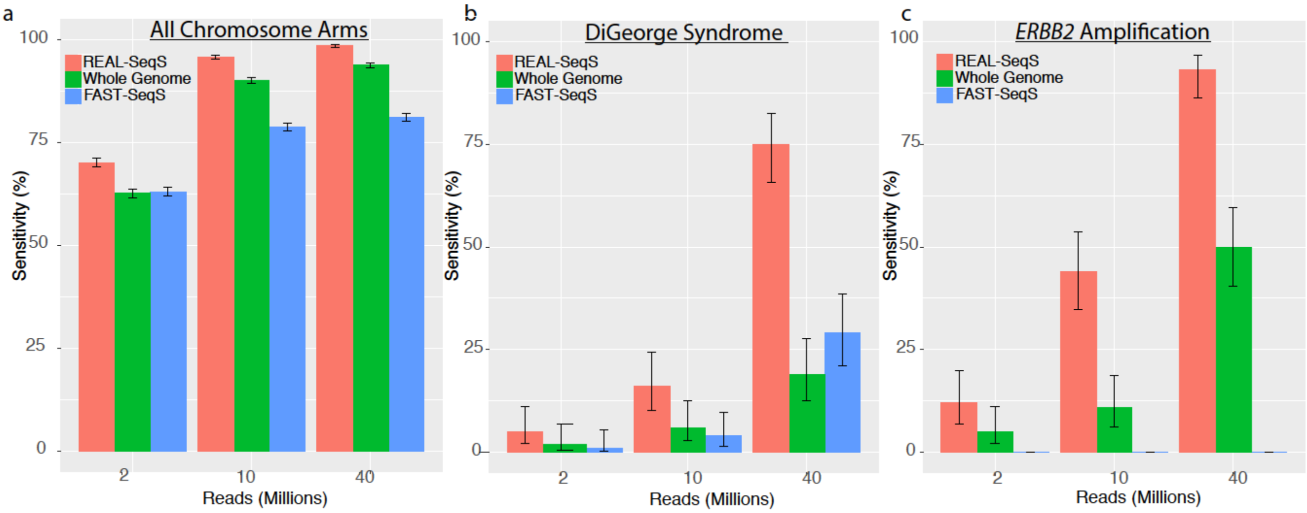

Figure 1. Detection of Aneuploidy using Next Generation Sequencing Technologies. Sensitivities were calculated using a threshold at 99% specificity. Error bars represent 95% confidence intervals. (a) Comparison of sensitivity for monosomies and trisomies across all 39 non-acrocentric chromosome arms at 5% cell fraction. (b) Comparison of sensitivity for the 1.5 Mb DiGeorge deletion on 22q at 5% cell fraction. (c) Comparison of sensitivity for a 20 copy *ERBB2* focal amplification at 1% cell fraction.

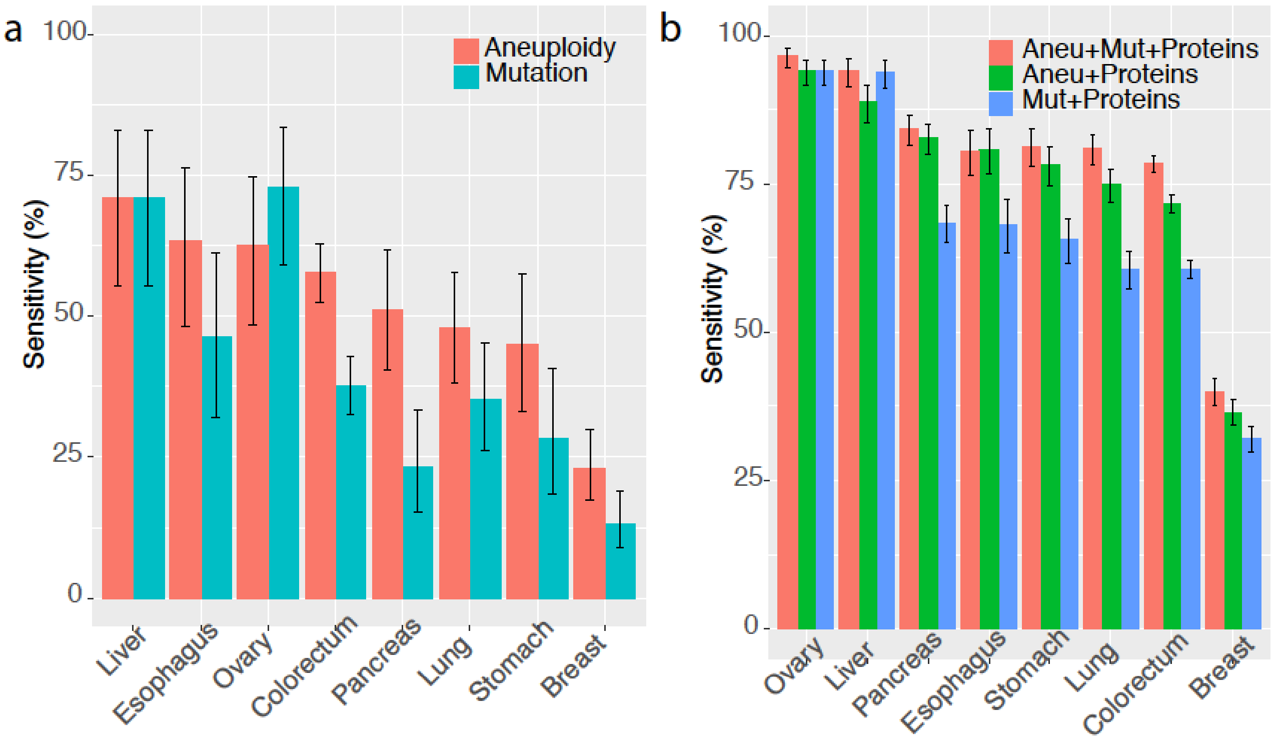

Figure 2. Detection of cancer in liquid biopsies from samples with non-metastatic cancers of eight different types. Sensitivities were calculated using a threshold at 99% specificity. Error bars represent 95% confidence intervals. (a) Comparison of aneuploidy status as calculated by REAL-SeqS to somatic mutations status. (b) Comparison of different multi-analyte tests. Three different multi-analyte tests evaluated sensitivity with and without the inclusion of aneuploidy status with somatic mutation status and 8 standard protein biomarkers.