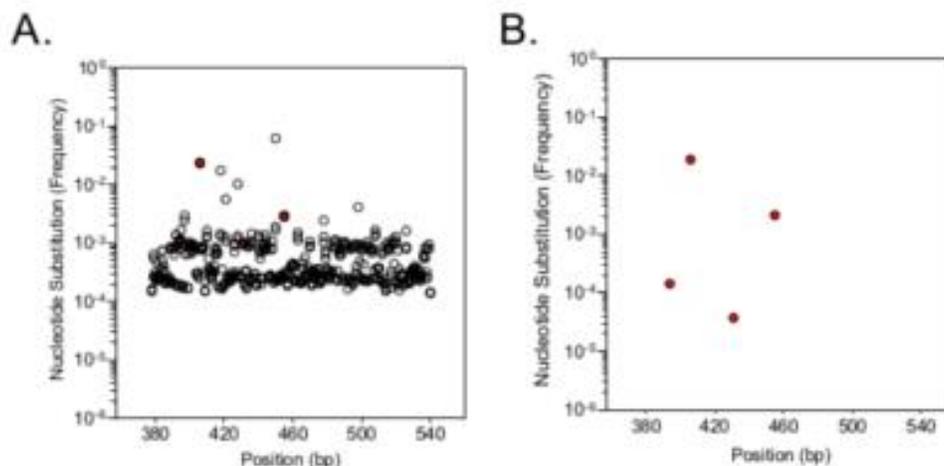


# A DNA sequencing sleuth called CypherSeq

November 15, 2015

A Neves



CypherSeq error correction and detection of rare mutations. A) Base substitutions determined by next-generation sequencing with a quality filter. The sample contains four mutations at the sites shown in red, which were spiked into the sample at known ratios of one mutant copy in 100, 1000, 10,000, and 100,000 wild type copies. However, due to the high background of erroneous substitutions, the real mutations cannot be distinguished from sequencing errors. B) Base substitutions determined by sequencing with CypherSeq. The increased accuracy of CypherSeq removes the background, allowing accurate and quantitative identification of the four real mutations down to a frequency of ~1 mutant in 100,000 copies.

*Image provided by Dr. Mark Gregory.*

Next-generation sequencing (NGS) technologies have transformed both basic and clinical research, as they have enabled the development of diverse applications ranging from pre-natal testing to cancer prognosis. However, there are a number of limitations that prevent NGS technologies from fulfilling their full potential. These include the inherent error of DNA sequencing instruments, the sequencing coverage depth and the amount of available sample DNA required. To address these limitations, post-doctoral fellows Drs. Mark Gregory and Jessica Bertout and collaborators in Dr. Jason Bielas' Laboratory (Human Biology and Public Health Sciences Divisions), developed a new method called CypherSeq that markedly improved NGS-based detection of rare mutations. This exciting report was recently published in the journal *Nucleic Acids Research*.

The investigators first designed the CypherSeq vector as a double-stranded barcoded vector that can be amplified in bacteria and contains all of the necessary sequences required to perform NGS. To delineate the power of the CypherSeq error correction algorithm, the researchers first applied the CypherSeq technology to a mixed DNA population with known ratios of both wild-type and four

mutant versions of the tumor suppressor gene *TP53*. This analysis revealed that a single *TP53* mutant could be resolved among more than  $10^5$  copies of wild-type *TP53*, which translates to a frequency of  $2.4 \times 10^{-7}$  mutations per base pair (see figure) and represents at least a 10-fold improvement over other NGS methods. Next, the authors addressed whether CypherSeq could be used for genome-wide assessment of both spontaneous and induced mutations. To this end, they grew liquid cultures of the baker's yeast *Saccharomyces cerevisiae* with or without the mutagen ethyl methanesulfonate (EMS). After application of CypherSeq sequencing and computational error-correction these samples revealed a base substitution frequency of  $1.4 \times 10^{-6}$  and  $4.6 \times 10^{-6}$  in untreated and EMS-treated cultures, respectively, demonstrating a clear increase due to treatment. Importantly, the additional mutations seen in the EMS-treated samples were nearly all C to T mutations, which match the known mutation signature of the EMS mutagen. As the EMS mutations in this experiment would only be present on a single genome within the yeast culture, the accurate detection of these mutations represent single cell resolution sequencing. Thus, CypherSeq is a more accurate DNA sequencing method that, for the first time, enables genome-wide calculations of mutations rates without resorting to single cell approaches.

Finally, the authors demonstrated that CypherSeq can also efficiently enrich a sequencing library for specific genomic sites. Because the CypherSeq vector is circular, rolling circle amplification (RCA) of a biotinylated primer was used to generate single-stranded DNA concatemers, which could be further enriched with streptavidin affinity purification. Indeed, CypherSeq vectors containing *TP53* sequences in the background of randomly sheared genomic DNA were enriched 977-fold. Such enrichment, decreases off-target superfluous sequencing, permits deeper sequencing depth of targeted sites, and thus the detection of much rarer variants. In summary, CypherSeq's highly sensitive error correction and RCA-based target enrichment strategies offer tantalizing possibilities for cancer diagnostics. Said Dr. Gregory: "Combining highly accurate sequencing with the ability to target specific regions of the genome is an important advance which will allow us to investigate rare mutations throughout the genome, at a resolution unmatched before. This technology should enable the development of a new class of sequencing-based clinical tests, such as noninvasive blood tests for early cancer diagnosis, identification of optimal therapeutic approaches, and monitoring of treatment response".

Funding for this study was provided by the National Institutes of Health, the Department of Defense, the Canary Foundation, the Marsha Rivkin Center for Ovarian Cancer Research, and the Ellison Medical Foundation.

[Gregory MT, Bertout JA, Ericson NG, Taylor SD, Mukherjee R, Robins HS, Drescher CW, Bielas JH.](#)

2015. Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic Acids Res.* doi: 10.1093/nar/gkv915. [Epub ahead of print.]