|  |
| --- |
| Category  Yan Guo1\*, Jiang Li1  1. Vanderbilt Ingram Cancer Center, Center for Quantitative Sciences, Nashville, TN  3. Center for Human Genetics Research  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Motivation:**

**Results:**

# introduction

The emergence and rapid maturity of next generation sequencing (NGS) technologies enabled massive high throughput production of sequencing data at low cost. Large NIH funded projects such as projects such as The Cancer Genome Atlas (TCGA) and the 1000 Genome project have generated huge amounts of sequencing data on tissues from a variety of diseased and healthy individuals. The data-storage database, Sequencing Read Archive (SRA) stores hundreds terabytes of raw sequencing data from the NGS platforms including Roche 454 GS System, Illumina Genome Analyzer, Applied Biosystems SOLiD System etc. Analysis on sequencing data has evolved rapidly over the last few years. However, majority of the analysis done on sequencing data were focused on targeted sequencing region from experiment design. NGS data is rich and informative and contains many off-target sequences that are often ignored, which may be biologically relevant. Previously, we have shown that sequencing data outside capture regions can produce reliable genetic variation data ([Guo, et al., 2012](#_ENREF_14)). In the same study, we showed that the capture efficiencies of commonly used capture kits (Illumina TrueSeq, Agilent SureSelect, and NimbleGen SeqCap EZ) are only between 40-60% efficient. A significant portion of DNA sequence reads aligned to regions outside the intended capture regions and thus is off-target. A recent study has shown that mitochondrial DNA sequences is sequenced in exome sequencing data ([Larman, et al., 2012](#_ENREF_18)), even when the mitochondrial DNA is not included in the target region. Based on those findings, we designed and implemented a tool, MitoSeek, for high throughput secondary mitochondria data mining from exome, whole genome sequencing data.

MitoSeek extracts mitochondria information from exome or whole genome sequencing data and computes four major mitochondria dysfunctions: heteroplasmy mutation, somatic mutation, relative copy number, and large structure change. Those four types of dysfunctions have all been linked to carcinogenesis or other mitochondria diseases. Warburg proposed a mechanism to explain the differences in energy metabolism between normal and cancer cells over half a century ago. Various studies have shown that mutations in mtDNA can contribute to cancer etiology ([Baysal, et al., 2000](#_ENREF_5); [Vanharanta, et al., 2004](#_ENREF_36)) and mtDNA mutations are associated with various types of cancer ([Canter, et al., 2005](#_ENREF_6); [Dasgupta, et al., 2008](#_ENREF_8); [Fliss, et al., 2000](#_ENREF_12); [Herrmann, et al., 2003](#_ENREF_15); [Petrosillo, et al., 2005](#_ENREF_28); [Sun, et al., 2009](#_ENREF_33)). Mammalian cells typically contain over 1,000 mitochondria, and each mitochondrion harbors 2-10 copies of mtDNA ([Robin and Wong, 1988](#_ENREF_30)). Thus, mtDNA mutations are usually heteroplasmic, with a mixture of mutant and wild-type mtDNA copies within a cell ([De Angelis, et al.](#_ENREF_9)). It has been shown that an mtDNA mutation does not need to reach homoplasmy, i.e., all copies of mtDNA within a cell are mutated, to promote tumor growth ([Lewis, et al., 2000](#_ENREF_19); [Park, et al., 2009](#_ENREF_26)). Researchers have found that both increased mtDNA copy number from whole blood DNA ([Shen, et al., 2010](#_ENREF_31)) and reduced mtDNA copy number in tissue may increase breast cancer risk ([Bai, et al., 2011](#_ENREF_3); [Tseng, et al., 2006](#_ENREF_35); [Yu, et al., 2007](#_ENREF_41)). Large-scale deletions in mitochondria occur less frequently than point mutations; nevertheless they contribute much more adverse effect to mitochondria related diseases. Deletions in mtDNA has been linked to aging ([Eshaghian, et al., 2006](#_ENREF_11); [Meissner, et al., 2008](#_ENREF_22)), cancer ([Abnet, et al., 2004](#_ENREF_1); [Kamalidehghan, et al., 2006](#_ENREF_16); [Tseng, et al., 2006](#_ENREF_35); [Wu, et al., 2005](#_ENREF_38)) and other diseases ([Degoul, et al., 1991](#_ENREF_10); [Manfredi, et al., 1997](#_ENREF_20); [Moraes, et al., 1989](#_ENREF_23); [Shoffner, et al., 1989](#_ENREF_32); [Zeviani, et al., 1988](#_ENREF_42)).

In this paper, we present MitoSeek, a novel tool designed for high throughput secondary data mining on mitochondria genome using exome sequencing or whole genome sequencing data. It is rich features and intuitive. In addition to the four mitochondria dysfunctions, MitoSeek offers many convenient features for conducting mitochondria research. We will describe the approach and algorithm behind each feature of MitoSeek in detail in the method section. In the result section, we demonstrate MitoSeek using real exome sequencing and whole genome sequencing data. MitoSeek and its source code can be downloaded at ???.

# method

## Overall Algorithm

Although the mitochondrial sequence is not explicitly targeted in the exome sequencing process, mitochondrial sequence information can be extracted from exome sequencing data. The input file to MitoSeek is BAM file from alignment result of exome or whole genome sequencing. Most NGS data repositories such as 1000 Genome Projects and TCGA offer direct download of BAM files. MitoSeek will parse the BAM file extract mitochondria information. Before conducting any analysis, MitoSeek will first produce mitochondria coverage quality control report which contains important statistics such as average depth, percent of base pairs covered, base quality distribution, mapping quality distribution and insert size distribution etc. Those quality control parameters serve as important indicator for the success of later analysis. We strongly recommend all users examine the quality control report before conducting any further analysis. MitoSeek is capable of detecting four types of mitochondria dysfunctions: heteroplasmy mutation, somatic mutation, relative copy, and large structural change.

## Heteroplasmy Detection

The single most crucial factor for detecting heteroplasmy mutation is depth. The most ideal sequencing technique for detecting heteroplasmy in mitochondria is mitochondria targeted sequencing which is capable of generating depth of up to 10 thousand, and detect heteroplamy as low as 0.1%. As byproduct, depth of mitochondria in exome sequencing and whole genome sequencing data are significantly lower (around hundreds) which limit the detectable heteroplasmy to around 1%. MitoSeek works with any sequencing data that contains mitochondria information. Based on the alignment quality control report, MitoSeek will automatic adjust heteroplasmy detection threshold to most appropriate level. Heteroplasmy detection threshold is defined on two scales: read count and read percentage. Read count denotes the number of reads we must observe to support heteroplasmy while read percentage denotes the percentage of reads we must observe to support heteroplasmy. Both scales can be used together or individually. The minimum recommended depth requirement for detecting heteroplasmy is 50. Lower depth will severely damage the confidence of heteroplasmy calling.

Result of heteroplasmy mutation is compiled in a comprehensive report which is composed of important information such as allele counts on forward and reverse strand, strand bias score, median mapping quality, confidence interval etc for each mitochondria position. Furthermore, heteroplasmy rate per sample is computed as heteroplasmy per 1000 base pair. MitoSeek also outputs circus plot for clear visualization of heteroplasmy mutation over mitochondria genes.

## Somatic Mutation Detection

Current genotype callers such as GATK’s Unified Genotyper ([McKenna, et al., 2010](#_ENREF_21)) developed by Broad Institute and glfMultiple developed by University of Michigan, are designed for diploid genome where heterozygous genotype is expected. Using those genotype callers on a haploid genome where only single allele is expected such as mitochondria will generate inaccurate results. The common way for identifying somatic mutation is to compare sequences between paired normal control and tumor samples. If we observe alternative alleles at a genomic position in the tumor but not in the matched normal control at the same position, we say this is an acquired somatic mutation. If we observe an alternative allele at a genomic position in the normal control but not in the tumor, we call it loss of heterozygosity (LOH). One common strategy used by many studies ([Nikolaev, et al., 2012](#_ENREF_25); [Vissers, et al., 2011](#_ENREF_37); [Yan, et al., 2011](#_ENREF_40)) to identify somatic mutations is to use the SNP caller to first determine the genotypes of tumor and control then compare the two. This type of approach has certain limitations. First of all, the threshold of mutation detection may be significantly different from a SNP. For example, for SNPs on diploid genomes, approximately 50 percent of the reads should support the alternative allele. However, for a somatic mutation, depending on the type of normal control samples used, the expected percentage of mutated reads might significantly differ. If blood sample is used as a control, we expect to observe germline mutations only, whereas if adjacent normal tissue is used as a control, the reads observed might represent a mixture of tumor and normal tissues due to tumor contamination, which can cause the SNP callers to make a false heterozygous inference. Furthermore, it’s possible that the tumor tissue sample is contaminated by the normal tissue. Even after purification procedures such as microdissection, the tumor percentage estimation is far from accurate. Lower tumor concentration in tumor sample might cause the sequencer to sequence insufficient amount of reads to support a heterozygous call by current SNP callers. To solve this problem, we propose to compare the empirical allele counts between tumor and normal control directly instead of using a genotype caller. By bypassing the inferring genotype, we can effectively detect small percentage of mutations that might otherwise not be detectable by a SNP caller and filter out potential wrong heterozygous inferences from normal controls. MitoSeek can extract empirical allele count for every mitochondria position then compare the allele counts between tumor and normal to determine somatic mutation status.

## Relative Copy Number Estimation

Recently, a study ([Picardi and Pesole, 2012](#_ENREF_29)) claimed that fraction of captured mitochondrial sequences is linkedto the relative abundance of the corresponding mitochondrialgenome in the original total DNA extract*.* Another group proved that mitochondria copy number can be obtained through NGS data ([Castle, et al., 2010](#_ENREF_7)). Based on those findings, we introduce a method for estimating relative mitochondria copy number through sequencing exome and whole sequencing data in MitoSeek. The method takes the advantage of the proportion of mitochondria reads captured during exome, and whole genome. The relative mitochondria copy number is computed as following: , where is the reads aligned to mitochondria and passed quality filter and is the total reads passed quality filter. Alternatively, relative mitochondria copy number can be computed as , where is the average depth of mitochondria, and is the average of exome. If whole genome data is used, will be replaced with , where is the average depth cross whole genome excluding mitochondria. The relative mitochondria copy number is useful when comparison between two conditions is needed.

## Structural Change

MitoSeek also reports several mitochondria structural change parameters when pair-end sequencing data is given as input. During alignment, a portion of the read-pairs will be discordantly mapped, meaning one read of the pair is aligned to mitochondria and the mate pair is aligned elsewhere. Such reads are like to be the results of alignment errors from homologous regions between mitochondria genome and other genomes. However, they could also indicate mitochondria integration into other genomes which has been reported to be possible by multiple studies ([Mourier, et al., 2001](#_ENREF_24); [Timmis, et al., 2004](#_ENREF_34)). Furthermore, MitoSeek can estimate the average DNA fragment size from the input BAM file. Significantly larger fragment size in mitochondria indicates large deletion. A summary of discordant read pairs and read pair with abnormal fragment size will be provided by MitoSeek.

## Other Features

The most common sequencing alignment reference for human is HG19 which has mitochondria genome at 16571 base pairs, however the most accepted mitochondria reference is the revised Cambridge Reference sequence (rCRS) ([Andrews, et al., 1999](#_ENREF_2)) (GenBank: NC\_012920) which has mitochondria genome at 16569 base pairs. In addition to the index difference, there are 40 nucleotide differences. MitoSeek can interchange genomic positions and reference nucleotide between HG19 and rCRS.

MitoSeek is designed with exome and whole genome sequencing in mind. However, it is also compatible with any sequencing data that contains mitochondria information including mitochondria targeted sequencing. For mitochondria targeted sequencing data, MitoSeek searches for an artifact that caused by over sequenced primer region. Affymetrix's Genechip Human Mitochondria Resequencing Array 2.0’s amplification kit is often used for capturing mitochondria DNA. The kit amplifies genomic DNA using PCR with two primer sets mito3 and mito1-2. The two primer sets generate 7814bp and 9307bp long fragments respectively. This protocol specifically amplifies the entire mitochondrial genome from genomic DNA using overlapping primers to eliminate bias that may be introduced if using PCR methods. However, the primers usually can’t be washed clean which causes over sequencing of the primer at the overlapping region. MitoSeek can automatically detect such overlapping primer regions and exclude them from analysis.

Due to the high mutability of mitochondria genome, a small amount positions will maybe tri-allelic or even four-allelic. Many of the tri-allelic and four-allelic are result of sequencing errors. However, there have been evidence supporting such multi-allele heteroplasmy’s existence from analysis of GenBank’s mitochondria genomes ([Pereira, et al., 2009](#_ENREF_27)) (table S). MitoSeek will report all positions with multi-allele heteroplasmy, and note its novelty based on previous findings from GeneBank.

MitoSeek computes many additional quality control parameters for heteroplasmy and somatic mutations. The most novel one is strand bias. When using Illumina high throughput short read data, sometimes the genotype inferred from the positive strand and negative strand are significantly different, with one homozygous and the other heterozygous. This phenomenon is known as strand bias. Sequencing data at a single position in the genome can be represented by a 2 by 2 table where a, c represent the forward and reverse strand allele counts of major allele, and b, d represent the forward and reverse strand allele counts of the minor allele. MitoSeek computes a strand bias score using the formula which has was introduced in a previous mitochondria sequencing study ([Guo, et al., 2012](#_ENREF_13)). Extremely high strand bias score (top 10%) is an indication for false positive detections. Another quality control parameter MitoSeek computes is the transition transversion (Ti/Tv) ratio which is often used as a quality control parameter for exome SNPs([Yan Guo, 2012](#_ENREF_39)). Ti/Tv ratio is around 3.0 for SNPs inside exons and about 2.0 elsewhere ([Bainbridge, et al., 2011](#_ENREF_4)). For mitochondria, studies have shown much higher preference to transition over transversion ([Guo, et al., 2012](#_ENREF_13); [Lanave, et al., 1986](#_ENREF_17); [Pereira, et al., 2009](#_ENREF_27)), and the ratio is reported anywhere from 21 to 38. Other statistics computed by MitoSeek that can be used to access the confidence of detected heteroplasmy mutation rate and somatic mutation rate includes median mapping quality and number of reads that support the mutation. Number of reads that support the mutation can be represent as 95 percent confidence interval of the mutation rate. It is computed as , where is the detected mutation rate, and is the total reads observed at this location. Higher number of reads will produce narrower confidence interval. This confidence interval is the most important measurement for the credibility of detected mutation. A mutation with 10% mutation rate observed from 10 reads has wider confidence interval comparing to a mutation observed from 100 reads.

Annotation of heteroplasmy and somatic mutations including gene names, positions amino acid changes are computed and reported by MitoSeek in both HG19 and rCRS coordinates.

# Result

We downloaded exome sequencing data on 8 breast cancer tumor normal paired samples from TCGA and tested MitoSeek using those data. The mitochondria quality control report can be seen in Figure.

Heteroplasmy mutation

Figure

Somatic Mutation

Figure

Copy Number

Table

Structural Change

Table or figure

# dISCUSSION

MitoSeek is the only sequencing analysis tool that report allele counts separate by forward and reverse strands.

**ACKNOWLEDGMENT**

**REFERENCES**

**Abnet, C.C.*, et al.* (2004) Control region mutations and the 'common deletion' are frequent in the mitochondrial DNA of patients with esophageal squamous cell carcinoma, *BMC Cancer*, 4, 30.**

**Andrews, R.M.*, et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, *Nat Genet*, 23, 147.**

**Bai, R.K.*, et al.* (2011) Mitochondrial DNA content varies with pathological characteristics of breast cancer, *J Oncol*, 2011, 496189.**

**Bainbridge, M.N.*, et al.* (2011) Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities, *Genome Biol*, 12, R68.**

**Baysal, B.E.*, et al.* (2000) Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma, *Science*, 287, 848-851.**

**Canter, J.A.*, et al.* (2005) Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African-American women, *Cancer Res*, 65, 8028-8033.**

**Castle, J.C.*, et al.* (2010) DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing, *BMC Genomics*, 11, 244.**

**Dasgupta, S.*, et al.* (2008) Mitochondrial cytochrome B gene mutation promotes tumor growth in bladder cancer, *Cancer Res*, 68, 700-706.**

**De Angelis, R.*, et al.* (2009) Breast cancer survivors in the United States: geographic variability and time trends, 2005-2015, *Cancer*, 115, 1954-1966.**

**Degoul, F.*, et al.* (1991) Different mechanisms inferred from sequences of human mitochondrial DNA deletions in ocular myopathies, *Nucleic Acids Res*, 19, 493-496.**

**Eshaghian, A.*, et al.* (2006) Mitochondrial DNA deletions serve as biomarkers of aging in the skin, but are typically absent in nonmelanoma skin cancers, *J Invest Dermatol*, 126, 336-344.**

**Fliss, M.S.*, et al.* (2000) Facile detection of mitochondrial DNA mutations in tumors and bodily fluids, *Science*, 287, 2017-2019.**

**Guo, Y.*, et al.* (2012) The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation, *Mutat Res*, 744, 154-160.**

**Guo, Y.*, et al.* (2012) Exome sequencing generates high quality data in non-target regions, *BMC genomics*, 13, 194.**

**Herrmann, P.C.*, et al.* (2003) Mitochondrial proteome: altered cytochrome c oxidase subunit levels in prostate cancer, *Proteomics*, 3, 1801-1810.**

**Kamalidehghan, B.*, et al.* (2006) Delta mtDNA4977 is more common in non-tumoral cells from gastric cancer sample, *Arch Med Res*, 37, 730-735.**

**Lanave, C.*, et al.* (1986) Transition and transversion rate in the evolution of animal mitochondrial DNA, *Biosystems*, 19, 273-283.**

**Larman, T.C.*, et al.* (2012) Spectrum of somatic mitochondrial mutations in five cancers, *Proc Natl Acad Sci U S A*, 109, 14087-14091.**

**Lewis, P.D.*, et al.* (2000) Detection of damage to the mitochondrial genome in the oncocytic cells of Warthin's tumour, *J Pathol*, 191, 274-281.**

**Manfredi, G.*, et al.* (1997) Association of myopathy with large-scale mitochondrial DNA duplications and deletions: which is pathogenic?, *Ann Neurol*, 42, 180-188.**

**McKenna, A.*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome research*, 20, 1297-1303.**

**Meissner, C.*, et al.* (2008) The 4977 bp deletion of mitochondrial DNA in human skeletal muscle, heart and different areas of the brain: a useful biomarker or more?, *Exp Gerontol*, 43, 645-652.**

**Moraes, C.T.*, et al.* (1989) Mitochondrial DNA deletions in progressive external ophthalmoplegia and Kearns-Sayre syndrome, *N Engl J Med*, 320, 1293-1299.**

**Mourier, T.*, et al.* (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus, *Mol Biol Evol*, 18, 1833-1837.**

**Nikolaev, S.I.*, et al.* (2012) Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma, *Nat Genet*, 44, 133-139.**

**Park, J.S.*, et al.* (2009) A heteroplasmic, not homoplasmic, mitochondrial DNA mutation promotes tumorigenesis via alteration in reactive oxygen species generation and apoptosis, *Hum Mol Genet*, 18, 1578-1589.**

**Pereira, L.*, et al.* (2009) The diversity present in 5140 human mitochondrial genomes, *Am J Hum Genet*, 84, 628-640.**

**Petrosillo, G.*, et al.* (2005) Mitochondrial dysfunction associated with cardiac ischemia/reperfusion can be attenuated by oxygen tension control. Role of oxygen-free radicals and cardiolipin, *Biochim Biophys Acta*, 1710, 78-86.**

**Picardi, E. and Pesole, G. (2012) Mitochondrial genomes gleaned from human whole-exome sequencing, *Nat Methods*, 9, 523-524.**

**Robin, E.D. and Wong, R. (1988) Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells, *J Cell Physiol*, 136, 507-513.**

**Shen, J.*, et al.* (2010) Mitochondrial copy number and risk of breast cancer: a pilot study, *Mitochondrion*, 10, 62-68.**

**Shoffner, J.M.*, et al.* (1989) Spontaneous Kearns-Sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial DNA deletion: a slip-replication model and metabolic therapy, *Proc Natl Acad Sci U S A*, 86, 7952-7956.**

**Sun, W.*, et al.* (2009) Mitochondrial mutations contribute to HIF1alpha accumulation via increased reactive oxygen species and up-regulated pyruvate dehydrogenease kinase 2 in head and neck squamous cell carcinoma, *Clin Cancer Res*, 15, 476-484.**

**Timmis, J.N.*, et al.* (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes, *Nat Rev Genet*, 5, 123-135.**

**Tseng, L.M.*, et al.* (2006) Mitochondrial DNA mutations and mitochondrial DNA depletion in breast cancer, *Genes Chromosomes Cancer*, 45, 629-638.**

**Vanharanta, S.*, et al.* (2004) Early-onset renal cell carcinoma as a novel extraparaganglial component of SDHB-associated heritable paraganglioma, *Am J Hum Genet*, 74, 153-159.**

**Vissers, L.E.*, et al.* (2011) Whole-exome sequencing detects somatic mutations of IDH1 in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA), *Am J Med Genet A*, 155A, 2609-2616.**

**Wu, C.W.*, et al.* (2005) Mitochondrial DNA mutations and mitochondrial DNA depletion in gastric cancer, *Genes Chromosomes Cancer*, 44, 19-28.**

**Yan Guo, J.L., Jing He, Chung-I Li, Qiuyin Cai, Xiao-Ou Shu, Wei Zheng, Chun Li (2012) Exome Sequencing Generates High Quality Data in Non-Target Regions, *BMC Genomics*.**

**Yan, X.J.*, et al.* (2011) Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia, *Nat Genet*, 43, 309-315.**

**Yu, M.*, et al.* (2007) Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients, *IUBMB Life*, 59, 450-457.**

**Zeviani, M.*, et al.* (1988) Deletions of mitochondrial DNA in Kearns-Sayre syndrome, *Neurology*, 38, 1339-1346.**

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)