# AnthOligo: Automating the design of oligonucleotides for capture/enrichment technologies

Pushkala Jayaraman[1], Timothy Mosbruger [1], Taishan Hu[1], Nikolaos G Tairis[1], Chao Wu[1], Peter M Clark[5], Monica D'Arcy[1], Deborah Ferriola[1], Katarzyna Mackiewicz[1], Xiaowu Gai[2], Dimitrios Monos[1,3], Mahdi Sarmady[1,3]

[1] Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA USA

[2] Division of Biomedical & Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA USA

[3] Perelman School of Medicine, University of Pennsylvania. Philadelphia, PA USA

Correspondence to: Dimitrios Monos (monosd@chop.edu) or Mahdi Sarmady (sarmadym@chop.edu)

## Abstract

**Summary**

A number of methods have been devised to address the need for targeted genomic resequencing. One of these methods, Region-specific extraction (RSE) of DNA is characterized by the capture of long DNA fragments (15-20 kb) by magnetic beads, after enzymatic extension of oligonucleotides hybridized to selected genomic regions. Facilitating the selection of the most optimal capture oligos targeting a region of interest, satisfying the properties of temperature (Tm) and entropy (ΔG), while minimizing the formation of primer dimers in a pooled experiment is therefore necessary. Manual design and selection of oligos becomes an extremely arduous task complicated by factors such as length of the target region and number of targeted regions. Here we describe, AnthOligo, a web-based application developed to optimally automate the process of

generation of oligo sequences to be used for the targeting and capturing the continuum of large and complex genomic regions. Apart from generating oligos for RSE, this program may have wider applications in the design of customizable internal oligos to be used as baits for gene panel analysis or even probes for large-scale comparative genomic hybridization (CGH) array processes.

**Implementation and Availability**

The application written in Java8 and run on Tomcat9 is a lightweight Java Spring MVC framework that provides the user with a simple interface to upload an input file in BED format and customize parameters for each task. A Redis-like *MapReduce* framework is implemented to run sub-tasks in parallel to optimize time and system resources alongside a 'task-queuing' system that runs submitted jobs as a server-side background daemon. The task of probe design in AnthOligo commences when a user uploads an input file and concludes with the generation of a result-set containing an optimal set of region-specific oligos.

AnthOligo is currently available as a public web application with URL: http://antholigo.chop.edu.

**KEYWORDS: region-specific extraction, oligo, primer design, enrichment, next-generation sequencing**

# Introduction

Massively parallel sequencing, in particular, short-read technologies such as Exome Sequencing have become important milestones in genomic diagnosis. Newer technologies[1-3], such as long-

read sequencing using linked-read strategy from 10x genomics[4] and single-molecule real-time (SMRT) sequencing approach from PacBio[5] focus on improving coverage over complex genomic regions to achieve finer resolution over sequence and structural rearrangements. Combining such a sequencing approach with a low-cost targeted enrichment methodology provides significant benefits of economy, data management and analysis and generates a resultant "capture" data that is further enriched for one's regions due to longer reads spanning gaps and complex repeat regions.

Region-specific extraction (RSE) of DNA is a solution-based technique for enrichment of defined genomic regions of interest. The method's cost-effective target-enrichment approach allows longer sequence templates up to 20 kb and a uniform depth of coverage across a region of interest.

Probe design for targeted enrichment is a requirement for any NGS test development. Although there exist many stand-alone tools and web-applications to help address requirements for varied target enrichment approaches, none can be implemented directly for the RSE method[8-17]. The advantage of this specific oligonucleotide design method is the ability to "space" the oligos evenly at a certain distance (thousands of bases) and thus achieve equivalent target specificity with fewer probes required as compared to the tiling approach (1X or 2X tiling density) by many custom "kit" provisions. Prior to automation of oligonucleotide design for capture/enrichment, an analyst would have to painstakingly filter the oligonucleotides to create sets of oligos manually by scanning a large matrix of dimer-dimer interactions. The task could exponentially increase in complexity and time when factors such as target region, size or number of regions increased. By streamlining the process of oligo design via an automated, statistically-motivated downstream processing algorithm [9, 14, 16], we estimate the tool saves man-hours by at least

10-fold. Here, we present AnthOligo, an automated application to design evenly-spaced capture oligos when provided with coordinates for genome-specific regions of interest. We have successfully implemented AnthOligo to design optimal capture oligos for the Zebrafish genomes[7] and additionally targeted and captured 4 MB section of the highly complex, MHC region in the human genome[6] in a solution-based capture. Most recently, additional sets of oligos have been designed, enriching the MHC by including publicly available MHC reference sequences from other cell lines that were either, partially known or fully completed [18]. The newest set of oligos have been successfully used in our new study (Manuscript in preparation).

# Implementation

**Step 1:** A region in the input file can range from a single exon to multiple megabases. A sliding window approach spanning 2kb overlapping every 100bp ensured thorough coverage of the region (Figure 1). Primer3[19] was used to generate internal oligos within each window using a repeat-masked reference sequence[20]. UCSC BLAT[21] was used to inspect sequence specificity across the oligos at a percentage identity threshold customized at 95%. The 'susceptibility' to form hairpins and duplexes was estimated by measuring their Tm and ΔG predictions by MFold[22] and UNAFold[23] for dimer stability based on the parameters of SantaLucia et al.[8, 10, 24, 25].

**Step 2:** For each region of interest, oligos that passed applicable thresholds from **Step1** were considered "candidates". The algorithm modeled the storage of oligos and specific properties like 'dimer interactions' and 'association by distance' in a directed acyclic graph(DAG)[16] (Figure 1). For RSE method to be able to capture the entire region of interest (ROI), the first few "seed" oligos must lie within a short window across the start of the region. The graph object

consisted of seed oligos or 'root nodes' and associated oligos became 'child nodes'. Each 'edge' represented the user-defined distance between the root and child nodes. A depth-first-search (DFS) was then carried out to walk through "completed paths" in each graph. A path was "complete" when the "leaf" oligo was found within the end of the target region. Each completed path formed a "set" of oligos for the given region.

**Step 3:** Design of optimal collection of oligos for target capture using multiplex PCR required combinatorial optimization solutions[11, 17] (Figure 1). The number of heterodimer combinations C for *n* oligos for each input region could be calculated as*:*

$$num\ combinations = nC^2$$

In order to get a resultant "*combination of set of oligos*" across all of the user-provided input regions, region-specific oligo sets were cross-compared across the input regions to ensure that oligos across regions did not dimerize with each other in solution. Every *m, k, p* number of oligos across M, K, P additional input regions increased this number of combinations somewhat exponentially:

$$num\ combinations = mC^2 + kC^2 + pC^2 + nC^2$$

With increasing region size and number of regions, this became computationally intensive akin to the Np-complete 'knapsack problem'. Heuristic optimization allowed for scalability without sacrificing quality of the capture design by returning the first available combination of oligo sets that satisfied our thresholds.

# Results and Discussion

Besides the published work (6,7), oligos have been designed for capturing several genomic

regions associated with Noonan Syndrome (8 genes), Type 1 Diabetes (9 genomic regions), Crohn's Disease (10 genomic regions) and retinitis pigmentosa (37 genomic regions) (available upon request). In each case the oligonucleotides performed well as observed by uniformity, sensitivity and average depth of coverage[6, 7]. To additionally validate the tool, the MHC of a random sample was captured and sequenced on the Illumina MiSeq. Alignment was performed using COX as reference, since the sample showed a closer match to COX than PGF reference. The average depth of coverage was estimated at 100x with 98.4% of positions >20X [Supplementary data Fig 1]. The reason we attempted another capture of the MHC region, besides the one published earlier (6), is because we needed to assess the success of the design using a random sample with unknown MHC sequence. The previously published capture (6) involved the PGF cell line, which has a known MHC sequence and the oligos were designed based on this known reference sequence. This time the Antholigo using a number of different reference MHC sequences (18) was used to generate a new set of oligos that presumably can target the MHC of any random DNA sample.

To capture sequence with acceptable range of accuracy and uniform representation across all the regions in multiplexed reactions, oligonucleotides must meet certain specifications in terms of sequence specificity, efficient oligo design with minimal interaction between the probes and optimal process time[14, 16, 17, 26]. AnthOligo was implemented to satisfy these requirements with the RSE method. It is well-understood that target capture design for multiplexed reactions is an NP-complete problem [14, 27]. Heuristic optimization was necessary to process large regions, upwards of 1Mb while identifying sets of evenly spaced capture oligonucleotides throughout the target region with target specificity[28]. Combinatorial approaches along with *MapReduce*

framework helped multi-thread memory-intensive and data-intensive tasks to run within an optimal time.

Sequence specificity is governed by multiple factors, the majority of which are repeats in the genome and the presence of pseudogenes [10, 11, 29-32]. AnthOligo's use of hard-masked reference file for generating oligos resolves this by avoiding possible repeat regions in the sequence. BLAT results were filtered by focusing on the specificity of the 3' subsequence[33].

Although AnthOligo was developed to support the RSE method, its current abilities and flexibility for future enhancements may have wider applications in designing internal oligos that can be used to target the MHC using CRISPR-Cas9, baits for gene panel analysis or even probes for CgH array processes. AnthOligo is thus, a unique tool to an unaddressed domain and results show that it achieves the desired objectives.

## Acknowledgements

Thanks to Juan Carlos Perin for a great name, Dr. Kajia Cao, Dr. Chao Wu for help with the NP-Complete optimization problem.

## Funding

The project described was supported by Award Number P30DK019525 from the National Institute of Diabetes and Digestive and Kidney Diseases to DM.

## References

1. Gnirke, A., et al., *Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.* Nat Biotechnol, 2009. **27**(2): p. 182-9.
2. Okou, D.T., et al., *Microarray-based genomic selection for high-throughput resequencing.* Nat Methods, 2007. **4**(11): p. 907-9.

3. Tewhey, R., et al., *Enrichment of sequencing targets from the human genome by solution hybridization.* Genome Biol, 2009. **10**(10): p. R116.
4. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.* Nature Biotechnology, 2019.
5. Zheng, G.X.Y., et al., *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing.* Nature Biotechnology, 2016. **34**: p. 303.
6. Dapprich, J., et al., *The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity.* BMC Genomics, 2016. **17**: p. 486.
7. Gupta, T., et al., *Microtubule actin crosslinking factor 1 regulates the Balbiani body and animal-vegetal polarity of the zebrafish oocyte.* PLoS Genet, 2010. **6**(8): p. e1001073.
8. Rouillard, J.-M., M. Zuker, and E. Gulari, *OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach.* Nucleic acids research, 2003. **31**(12): p. 3057-3062.
9. Ben Zakour, N., et al., *GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification.* Nucleic Acids Res, 2004. **32**(1): p. 17-24.
10. Vallone, P.M. and J.M. Butler, *AutoDimer: a screening tool for primer-dimer and hairpin structures.* Biotechniques, 2004. **37**(2): p. 226-31.
11. Nordberg, E.K., *YODA: selecting signature oligonucleotides.* Bioinformatics, 2005. **21**(8): p. 1365-70.
12. Jabado, O.J., et al., *Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments.* Nucleic Acids Res, 2006. **34**(22): p. 6605-11.
13. Rychlik, W., *OLIGO 7 primer analysis software.* Methods Mol Biol, 2007. **402**: p. 35-60.
14. Shen, Z., et al., *MPprimer: a program for reliable multiplex PCR primer design.* BMC Bioinformatics, 2010. **11**: p. 143.
15. Ilie, L., et al., *BOND: Basic OligoNucleotide Design.* BMC Bioinformatics, 2013. **14**: p. 69.
16. Francis, F., M.D. Dumas, and R.J. Wisser, *ThermoAlign: a genome-aware primer design tool for tiled amplicon resequencing.* Sci Rep, 2017. **7**: p. 44437.
17. Wingo, T.S., A. Kotlar, and D.J. Cutler, *MPD: multiplex primer design for next-generation targeted sequencing.* BMC Bioinformatics, 2017. **18**(1): p. 14.
18. Horton, R., et al., *Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project.* Immunogenetics, 2008. **60**(1): p. 1-18.
19. Untergasser, A., et al., *Primer3--new capabilities and interfaces.* Nucleic Acids Res, 2012. **40**(15): p. e115.
20. Chen, N., *Using RepeatMasker to identify repetitive elements in genomic sequences.* Curr Protoc Bioinformatics, 2004. **Chapter 4**: p. Unit 4 10.
21. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
22. Zuker, M., D.H. Mathews, and D.H. Turner, *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide*, in *RNA Biochemistry and Biotechnology*, J. Barciszewski and B.F.C. Clark, Editors. 1999, Springer Netherlands: Dordrecht. p. 11-43.

23. Markham, N.R. and M. Zuker, *UNAFold: software for nucleic acid folding and hybridization.* Methods Mol Biol, 2008. **453**: p. 3-31.
24. SantaLucia, J., Jr., *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.* Proc Natl Acad Sci U S A, 1998. **95**(4): p. 1460-5.
25. Owczarzy, R., et al., *IDT SciTools: a suite for analysis and design of nucleic acid oligomers.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W163-9.
26. Mulle, J.G., et al., *Empirical evaluation of oligonucleotide probe selection for DNA microarrays.* PLoS One, 2010. **5**(3): p. e9921.
27. Nicodeme, P. and J.M. Steyaert, *Selecting optimal oligonucleotide primers for multiplex PCR.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 210-3.
28. Hysom, D.A., et al., *Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments.* PLOS ONE, 2012. **7**(4): p. e34560.
29. Claes, K.B. and K. De Leeneer, *Dealing with pseudogenes in molecular diagnostics in the next-generation sequencing era.* Methods Mol Biol, 2014. **1167**: p. 303-15.
30. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2011. **13**(1): p. 36-46.
31. Mertes, F., et al., *Targeted enrichment of genomic DNA regions for next-generation sequencing.* Brief Funct Genomics, 2011. **10**(6): p. 374-86.
32. Koressaar, T., et al., *Primer3_masker: integrating masking of template sequence with primer design software.* Bioinformatics, 2018. **34**(11): p. 1937-1938.
33. Miura, F., et al., *A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences.* Bioinformatics, 2005. **21**(24): p. 4363-70.