

1 AnthOligo: Automating the design of oligonucleotides

2 for capture/enrichment technologies

3 Pushkala Jayaraman¹, Timothy Mosbruger¹, Taishan Hu¹, Nikolaos G Tairis¹, Chao Wu¹, Peter
4 M Clark⁵, Monica D'Arcy¹, Deborah Ferriola¹, Katarzyna Mackiewicz¹, Xiaowu Gai², Dimitrios
5 Monos^{1,3}, Mahdi Sarmady^{1,3}

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

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

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

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

11

12 Abstract

13 Summary

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

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

28

29 **Implementation and Availability**

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

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

39

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

42

43 **Introduction**

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

46 read sequencing using linked-read strategy from 10x genomics[4] and single-molecule real-time
47 (SMRT) sequencing approach from PacBio[5] focus on improving coverage over complex
48 genomic regions to achieve finer resolution over sequence and structural rearrangements.
49 Combining such a sequencing approach with a low-cost targeted enrichment methodology
50 provides significant benefits of economy, data management and analysis and generates a
51 resultant “capture” data that is further enriched for one’s regions due to longer reads spanning
52 gaps and complex repeat regions.
53 Region-specific extraction (RSE) of DNA is a solution-based technique for enrichment of
54 defined genomic regions of interest. The method’s cost-effective target-enrichment approach
55 allows longer sequence templates up to 20 kb and a uniform depth of coverage across a region of
56 interest.

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

69 10-fold. Here, we present AnthOligo, an automated application to design evenly-spaced capture
70 oligos when provided with coordinates for genome-specific regions of interest. We have
71 successfully implemented AnthOligo to design optimal capture oligos for the Zebrafish
72 genomes[[7](#)] and additionally targeted and captured 4 MB section of the highly complex, MHC
73 region in the human genome[[6](#)] in a solution-based capture. Most recently, additional sets of
74 oligos have been designed, enriching the MHC by including publicly available MHC reference
75 sequences from other cell lines that were either, partially known or fully completed [[18](#)]. The
76 newest set of oligos have been successfully used in our new study (Manuscript in preparation).

77

78 **Implementation**

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

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

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

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

$$\text{num combinations} = nC^2$$

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

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

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

109 **Results and Discussion**

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

111 regions associated with Noonan Syndrome (8 genes), Type 1 Diabetes (9 genomic regions),
112 Crohn's Disease (10 genomic regions) and retinitis pigmentosa (37 genomic regions) (available
113 upon request). In each case the oligonucleotides performed well as observed by uniformity,
114 sensitivity and average depth of coverage[[6](#), [7](#)]. To additionally validate the tool, the MHC of a
115 random sample was captured and sequenced on the Illumina MiSeq. Alignment was performed
116 using COX as reference, since the sample showed a closer match to COX than PGF reference.
117 The average depth of coverage was estimated at 100x with 98.4% of positions >20X
118 [Supplementary data Fig 1]. The reason we attempted another capture of the MHC region,
119 besides the one published earlier (6), is because we needed to assess the success of the design
120 using a random sample with unknown MHC sequence. The previously published capture (6)
121 involved the PGF cell line, which has a known MHC sequence and the oligos were designed
122 based on this known reference sequence. This time the Antholigo using a number of different
123 reference MHC sequences (18) was used to generate a new set of oligos that presumably can
124 target the MHC of any random DNA sample.
125 To capture sequence with acceptable range of accuracy and uniform representation across all the
126 regions in multiplexed reactions, oligonucleotides must meet certain specifications in terms of
127 sequence specificity, efficient oligo design with minimal interaction between the probes and
128 optimal process time[[14](#), [16](#), [17](#), [26](#)]. AnthOligo was implemented to satisfy these requirements
129 with the RSE method. It is well-understood that target capture design for multiplexed reactions is
130 an NP-complete problem [[14](#), [27](#)]. Heuristic optimization was necessary to process large regions,
131 upwards of 1Mb while identifying sets of evenly spaced capture oligonucleotides throughout the
132 target region with target specificity[[28](#)]. Combinatorial approaches along with *MapReduce*

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

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

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

144

145 **Acknowledgements**

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

148 **Funding**

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

151 **References**

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

156 3. Tewhey, R., et al., *Enrichment of sequencing targets from the human genome by solution*
157 *hybridization*. *Genome Biol*, 2009. **10**(10): p. R116.

158 4. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant*
159 *detection and assembly of a human genome*. *Nature Biotechnology*, 2019.

160 5. Zheng, G.X.Y., et al., *Haplotyping germline and cancer genomes with high-throughput*
161 *linked-read sequencing*. *Nature Biotechnology*, 2016. **34**: p. 303.

162 6. Dapprich, J., et al., *The next generation of target capture technologies - large DNA*
163 *fragment enrichment and sequencing determines regional genomic variation of high*
164 *complexity*. *BMC Genomics*, 2016. **17**: p. 486.

165 7. Gupta, T., et al., *Microtubule actin crosslinking factor 1 regulates the Balbiani body and*
166 *animal-vegetal polarity of the zebrafish oocyte*. *PLoS Genet*, 2010. **6**(8): p. e1001073.

167 8. Rouillard, J.-M., M. Zuker, and E. Gulari, *OligoArray 2.0: design of oligonucleotide probes*
168 *for DNA microarrays using a thermodynamic approach*. *Nucleic acids research*, 2003.
169 **31**(12): p. 3057-3062.

170 9. Ben Zakour, N., et al., *GenoFrag: software to design primers optimized for whole*
171 *genome scanning by long-range PCR amplification*. *Nucleic Acids Res*, 2004. **32**(1): p. 17-
172 24.

173 10. Vallone, P.M. and J.M. Butler, *AutoDimer: a screening tool for primer-dimer and hairpin*
174 *structures*. *Biotechniques*, 2004. **37**(2): p. 226-31.

175 11. Nordberg, E.K., *YODA: selecting signature oligonucleotides*. *Bioinformatics*, 2005. **21**(8):
176 p. 1365-70.

177 12. Jabado, O.J., et al., *Greene SCPrimer: a rapid comprehensive tool for designing*
178 *degenerate primers from multiple sequence alignments*. *Nucleic Acids Res*, 2006. **34**(22):
179 p. 6605-11.

180 13. Rychlik, W., *OLIGO 7 primer analysis software*. *Methods Mol Biol*, 2007. **402**: p. 35-60.

181 14. Shen, Z., et al., *MPprimer: a program for reliable multiplex PCR primer design*. *BMC*
182 *Bioinformatics*, 2010. **11**: p. 143.

183 15. Ilie, L., et al., *BOND: Basic OligoNucleotide Design*. *BMC Bioinformatics*, 2013. **14**: p. 69.

184 16. Francis, F., M.D. Dumas, and R.J. Wisser, *ThermoAlign: a genome-aware primer design*
185 *tool for tiled amplicon resequencing*. *Sci Rep*, 2017. **7**: p. 44437.

186 17. Wingo, T.S., A. Kotlar, and D.J. Cutler, *MPD: multiplex primer design for next-generation*
187 *targeted sequencing*. *BMC Bioinformatics*, 2017. **18**(1): p. 14.

188 18. Horton, R., et al., *Variation analysis and gene annotation of eight MHC haplotypes: the*
189 *MHC Haplotype Project*. *Immunogenetics*, 2008. **60**(1): p. 1-18.

190 19. Untergasser, A., et al., *Primer3--new capabilities and interfaces*. *Nucleic Acids Res*, 2012.
191 **40**(15): p. e115.

192 20. Chen, N., *Using RepeatMasker to identify repetitive elements in genomic sequences*. *Curr*
193 *Protoc Bioinformatics*, 2004. **Chapter 4**: p. Unit 4 10.

194 21. Kent, W.J., *BLAT--the BLAST-like alignment tool*. *Genome Res*, 2002. **12**(4): p. 656-64.

195 22. Zuker, M., D.H. Mathews, and D.H. Turner, *Algorithms and Thermodynamics for RNA*
196 *Secondary Structure Prediction: A Practical Guide*, in *RNA Biochemistry and*
197 *Biotechnology*, J. Barciszewski and B.F.C. Clark, Editors. 1999, Springer Netherlands:
198 Dordrecht. p. 11-43.

199 23. Markham, N.R. and M. Zuker, *UNAFold: software for nucleic acid folding and*
200 *hybridization.* Methods Mol Biol, 2008. **453**: p. 3-31.

201 24. SantaLucia, J., Jr., *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-*
202 *neighbor thermodynamics.* Proc Natl Acad Sci U S A, 1998. **95**(4): p. 1460-5.

203 25. Owczarzy, R., et al., *IDT SciTools: a suite for analysis and design of nucleic acid*
204 *oligomers.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W163-9.

205 26. Mulle, J.G., et al., *Empirical evaluation of oligonucleotide probe selection for DNA*
206 *microarrays.* PLoS One, 2010. **5**(3): p. e9921.

207 27. Nicodeme, P. and J.M. Steyaert, *Selecting optimal oligonucleotide primers for multiplex*
208 *PCR.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 210-3.

209 28. Hysom, D.A., et al., *Skip the Alignment: Degenerate, Multiplex Primer and Probe Design*
210 *Using K-mer Matching Instead of Alignments.* PLOS ONE, 2012. **7**(4): p. e34560.

211 29. Claes, K.B. and K. De Leeneer, *Dealing with pseudogenes in molecular diagnostics in the*
212 *next-generation sequencing era.* Methods Mol Biol, 2014. **1167**: p. 303-15.

213 30. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing:*
214 *computational challenges and solutions.* Nat Rev Genet, 2011. **13**(1): p. 36-46.

215 31. Mertes, F., et al., *Targeted enrichment of genomic DNA regions for next-generation*
216 *sequencing.* Brief Funct Genomics, 2011. **10**(6): p. 374-86.

217 32. Koreasaar, T., et al., *Primer3_masker: integrating masking of template sequence with*
218 *primer design software.* Bioinformatics, 2018. **34**(11): p. 1937-1938.

219 33. Miura, F., et al., *A novel strategy to design highly specific PCR primers based on the*
220 *stability and uniqueness of 3'-end subsequences.* Bioinformatics, 2005. **21**(24): p. 4363-
221 70.

222

CREATE CANDIDATE OLIGOS FOR REGION
(detailed subworkflow)

