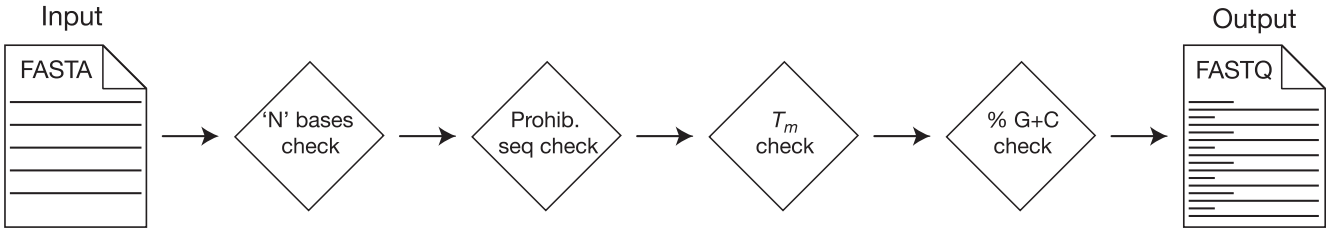
# Supporting Information

## Beliveau et al. 10.1073/pnas.1714530115

| Feature | OligoArray | OligoMiner |
|---|---|---|
| Specify probe length range | ✓ | ✓ |
| Specify GC content range | ✓ | ✓ |
| Specify probe $T_m$ range | ✓ | ✓ |
| Specify NN thermodynamic table | ✗ | ✓ |
| Specify sodium concentration | ✗ | ✓ |
| Specify formamide concentration | ✗ | ✓ |
| Specify probe concentration | ✗ | ✓ |
| Specify probe spacing | ✓ | ✓ |
| Choose alignment program & settings | ✗ | ✓ |
| Check for secondary structure | ✓ | ✓ |
| Check for high-abundance k-mers | ✗ | ✓ |
| Write log file | ✓ | ✓ |

**Fig. S1.** Description of features available in OligoArray and OligoMiner.

A *blockParse.py*

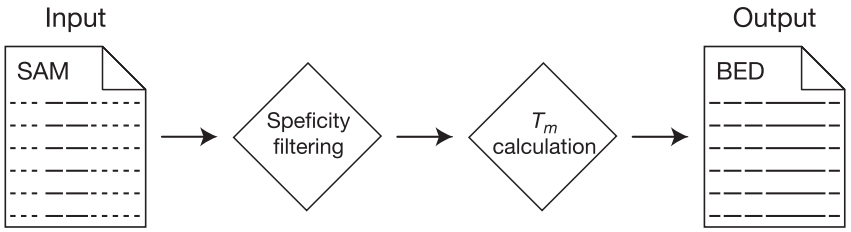



B

| Option | Default | Usage |
|---|---|---|
| -f /--file | None | **Required**. Specifies the FASTA-formatted file to find candidate probes in. |
| -l /--minLength | 36 | The minimum allowed probe length. |
| -L /--maxLength | 41 | The maximum allowed probe length. |
| -g /--min_GC | 20 | The minimum % G + C bases allowed. |
| -G /--max_GC | 80 | The maximum % G + C bases allowed. |
| -t /--min_Tm | 42 | The minimum salt + formamide adjusted $T_m$ allowed. |
| -T /--max_Tm | 47 | The maximum salt + formamide adjusted $T_m$ allowed. |
| -X /--prohibitedSeqs | 'AAAAA,TTTTT, CCCCC,GGGGG' | Prohibited sequence list (separated by commas with no spaces). Any candidate probe containing one of these sequences will be rejected. |
| -s /--salt | 390 | The mM $Na^+$ concentration, default is 390 to match 2X SSC. |
| -F /--formamide | 50 | The percent formamide being used. |
| -S /--Spacing | 0 | The minimum spacing (bp) between adjacent probes. |
| -c /--dnac1 | 25 | nM concentration of higher conc. strand (e.g. the probe) for thermo. calcs. |
| -C /--dnac2 | 25 | nM concentration of lower conc. strand (e.g. the target) for thermo. calcs. |
| -n /--nn_table | 'DNA_NN3' | The Biopython Nearest Neighbor thermodynamic table to use. |
| -H /--header | None | Specifies a custom FASTA header in the format chr:start-stop. |
| -b /--bed | False | Writes output as a BED file instead of a FASTQ file if flagged. |
| -O /--OverlapMode | False | Finds all possible candidate probes and ignores '-S' value if flagged. |
| -v /--verbose | False | Prints info about probe discovery progress as the script runs if flagged. |
| -R /--Report | False | Writes detailed log file about the behavior of the script if flagged. |
| -D /--Debug | False | Prints detailed info about the behavior of the script as it runs if flagged. |
| -M /--Meta | False | Writes a small summary file describing the outcome of the run if flagged. |
| -o /--output | False | Specifies the stem of the output filename. |

**Fig. S2.** Description of blockParse. (*A*) Schematic diagram illustrating the nature and order of checks used to screen candidate probe sequences. (*B*) Description of command-line options and the default values/settings for each.

A *outputClean.py*




B

| Option | Default | Usage |
|---|---|---|
| -f /--file | None | **Required**. Specifies the SAM file to process. |
| -l /--lda | True | Filter the SAM file using the LDA model. |
| -u /--unique | False | Filter using unique mode; only keep candidates with 1 reported alignments. |
| -0 /--zero | False | Filter using zero mode; only keep candidates with 0 reported alignments. |
| -p /--prob | 0.5 | The probability threshold for classifying a candidate probe as likely to have off-target binding using the LDA model. Selecting larger values will improve precision (fewer false positives), but at the expense of recall (more false negatives). Selecting lower values will improve recall at the expense of precision. |
| -T /--Temp | 42 | The temperature-specific LDA model to use ( 32, 37, 42, 47, 52, or 57). |
| -s /--salt | 390 | The mM $Na^+$ concentration, default is 390 to match 2X SSC. |
| -F /--formamide | 50 | The percent formamide being used. |
| -R /--Report | False | Writes detailed log file about the behavior of the script if flagged. |
| -D /--Debug | False | Prints detailed info about the behavior of the script as it runs if flagged. |
| -M /--Meta | False | Writes a small summary file describing the outcome of the run if flagged. |
| -o /--output | False | Specifies the stem of the output filename. |

**Fig. S3.** Description of outputClean. (*A*) Schematic diagram illustrating the task order of the script. (*B*) Description of command-line options and the default values/settings for each.

A *32°C LDA Model*

| Class | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| p(duplexing) < 0.2 | 0.84 | 0.96 | 0.89 | 52,830 |
| p(duplexing) ≥ 0.2 | 0.98 | 0.91 | 0.94 | 109,772 |
| Average/total: | 0.93 | 0.93 | 0.93 | 162,602 |

B *37°C LDA Model*

| Class | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| p(duplexing) < 0.2 | 0.86 | 0.97 | 0.91 | 66,706 |
| p(duplexing) ≥ 0.2 | 0.98 | 0.89 | 0.93 | 96,073 |
| Average/total: | 0.93 | 0.92 | 0.92 | 162,779 |

C *42°C LDA Model*

| Class | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| p(duplexing) < 0.2 | 0.91 | 0.97 | 0.94 | 85,640 |
| p(duplexing) ≥ 0.2 | 0.96 | 0.89 | 0.93 | 77,252 |
| Average/total: | 0.93 | 0.93 | 0.93 | 162,892 |

D *47°C LDA Model*

| Class | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| p(duplexing) < 0.2 | 0.95 | 0.95 | 0.95 | 109,068 |
| p(duplexing) ≥ 0.2 | 0.91 | 0.90 | 0.91 | 53,861 |
| Average/total: | 0.94 | 0.94 | 0.94 | 162,929 |

E *52°C LDA Model*

| Class | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| p(duplexing) < 0.2 | 0.96 | 0.96 | 0.96 | 135,249 |
| p(duplexing) ≥ 0.2 | 0.82 | 0.82 | 0.82 | 27,548 |
| Average/total: | 0.94 | 0.94 | 0.94 | 162,797 |

F *57°C LDA Model*

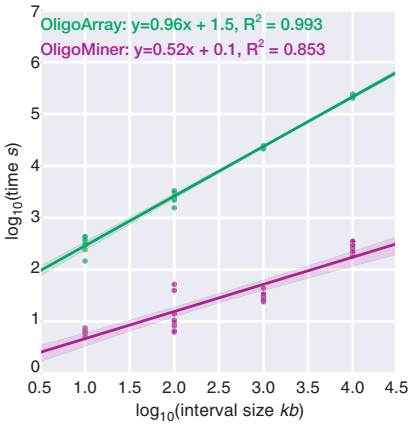
| Class | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| p(duplexing) < 0.2 | 0.97 | 0.99 | 0.98 | 156,633 |
| p(duplexing) ≥ 0.2 | 0.67 | 0.31 | 0.43 | 6,278 |
| Average/total: | 0.96 | 0.97 | 0.96 | 162,911 |

**Fig. S4.** Summary information for each temperature-specific LDA model. (*A–F*) For each temperature, the precision, recall, support-weighted $F_1$ score, and support are given.

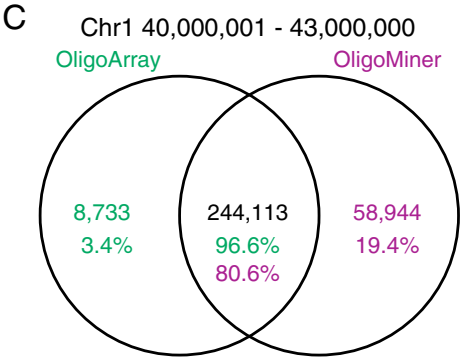
A

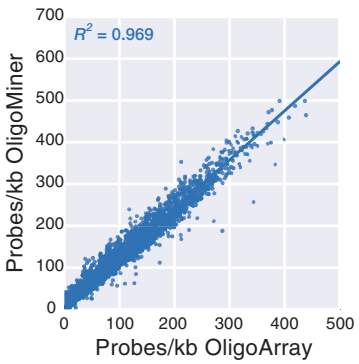
| Interval (kb) | Time OligoArray (s) | Time OligoMiner (s) | Fold Increase |
|---|---|---|---|
| 10 (*n* = 8) | 321 ± 96 | 6.3 ± 0.7 | 51x |
| 100 (*n* = 8) | 2,493 ± 547 | 18.4 ± 17 | 135x |
| 1,000 (*n* = 8) | 23,179 ± 1,200 | 30.4 ± 6.4 | 762x |
| 10,000 (*n* = 8) | 226,024 ± 12,610 | 271 ± 62 | 834x |

B

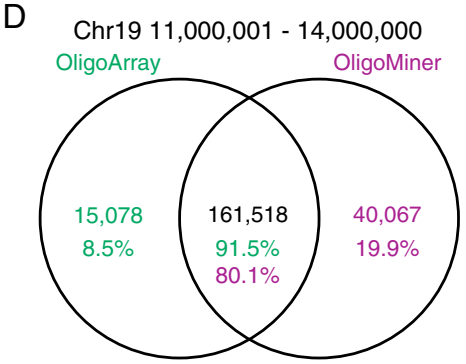


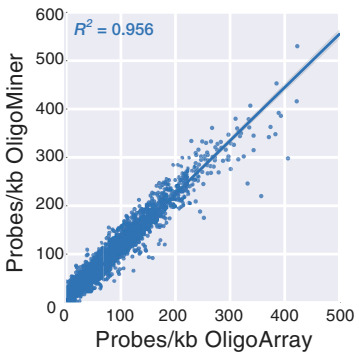

C



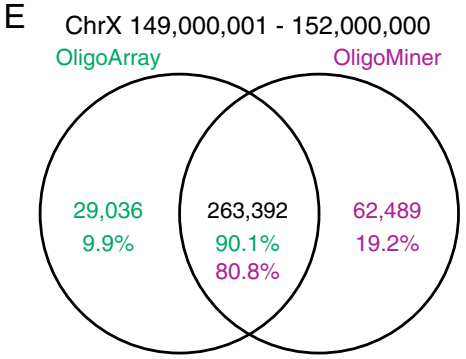




D




E



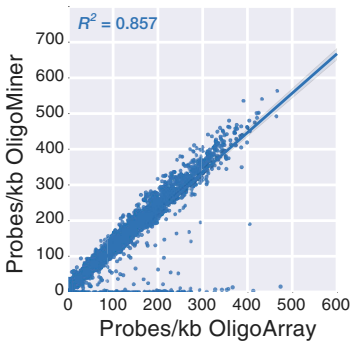




**Fig. S5.** Speed and coverage comparison between OligoArray and OligoMiner. (*A*) Mean times ± SD for probe discovery in eight 10-kb, 100-kb, 1-Mb, or 10-Mb intervals using OligoArray and OligoMiner. The fold increase in speed provided by OligoMiner for each interval is also shown. (*B*) Log-log plot of the data from *A*. Linear regression trend lines, equations, and $R^2$ values are also shown; shading represents the 95% CI of the fit. (*C–E*) Venn diagrams (*Left*) and plots showing linear regressions (*Right*) of three 3-Mb intervals mined exhaustively using OligoArray and OligoMiner. Plots show the number of probes discovered in each nonoverlapping 1-kb window starting with the first coordinate of each interval. Linear regression trend lines and $R^2$ values are also shown; shading represents the 95% CI of the fit.

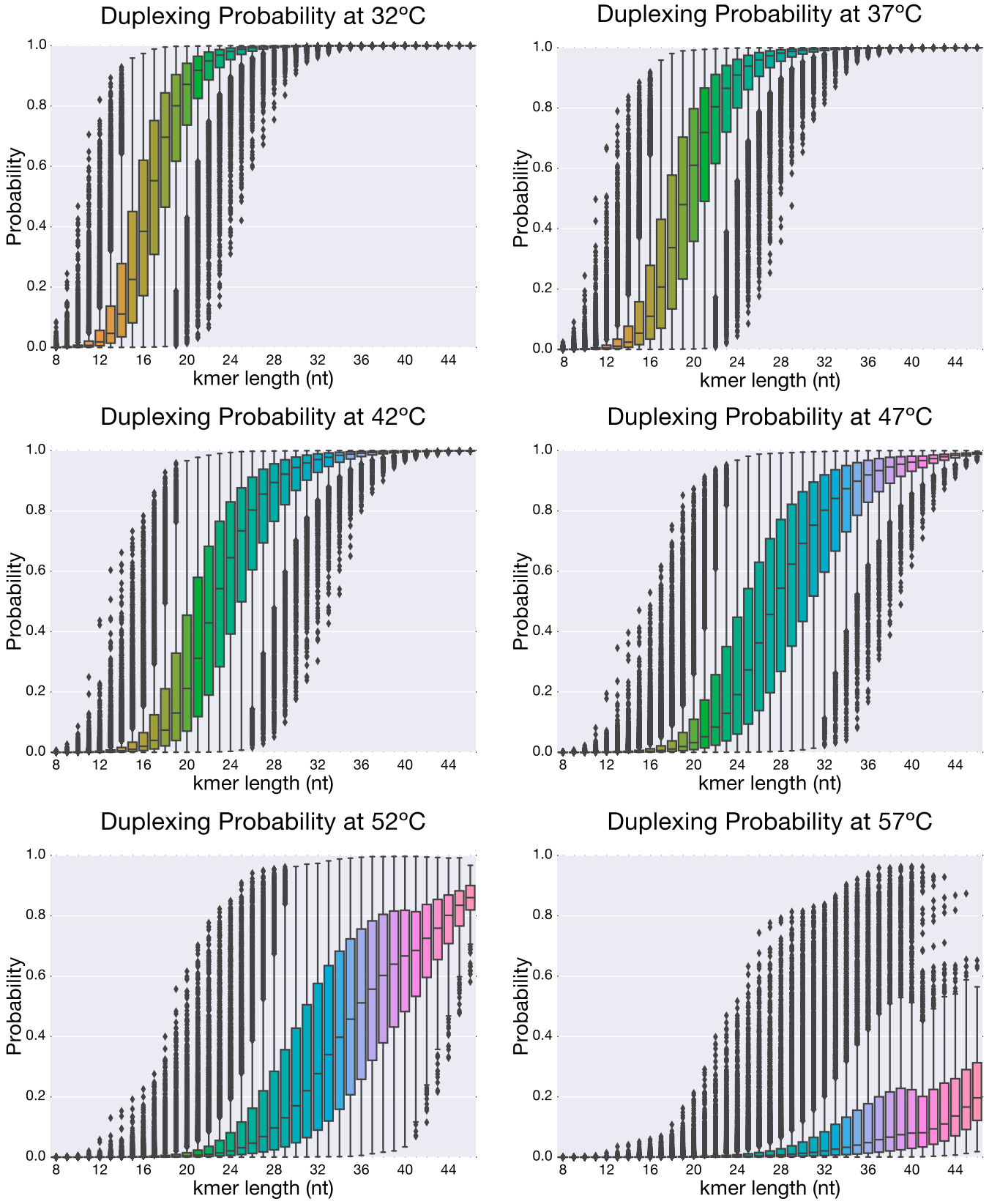



**Fig. S6.** Box plots depicting the duplexing probabilities of all kmers ≥8 nt in length with the reverse complements of their 40–46mer parental sequences at six different simulation temperatures in 390 mM $Na^+$ and 50% formamide.

**Table S1. Description of Oligopaint probe sets used**

| Probe set | Target | Size, kb | Complexity | Density, probes per kilobase | Mining parameters | Synthesis method |
|---|---|---|---|---|---|---|
| Xq28 "X.1" | hg38 chrX:149,681,511–151,005,677 | 1,324 | 5,415 | 4.1 | 40–45mers, 47–52 °C Tm; UM, no kmerFilter | Gel extraction |
| Xq28 "X.2" | hg38 chrX:151,015,314–153,048,806 | 2,033 | 6,482 | 3.2 | 40–45mers, 47–52 °C Tm; UM, no kmerFilter | Gel extraction |
| Xq28 "X.3" | hg38 chrX:153,051,165–153,868,151 | 817 | 4,776 | 5.8 | 40–45mers, 47–52 °C Tm; UM, no kmerFilter | Gel extraction |
| 19p13.2 "19.1" | hg19 chr19:9,921,047–11,719,207 | 1,798 | 10,985 | 6.1 | 36–41mers, 42–47 °C Tm; 47 °C LDM, kmerFilter -m 18 -k 5 | T7 |
| 19p13.2 "19.2" | hg19 chr19:11,725,035–12,760,026 | 1,035 | 3,678 | 3.6 | 36–41mers, 42–47 °C Tm; 47 °C LDM, kmerFilter -m 18 -k 5 | T7 |
| 19p13.2 20 kb | hg19 chr19:13,689,983–13,709,902 | 20 | 104 | 5.2 | 36–41mers, 42–47 °C Tm; 47 °C LDM, kmerFilter -m 18 -k 5 | T7 |
| Xist RNA | hg38 chrX:73,841,382–73,852,735 | 11 | 167 | 15.2 | 36–41mers, 42–47 °C Tm; 42 °C LDM, kmerFilter -m 18 -k 5 | Commercial column synthesis |

**Table S2. Description of oligo sequences used for probe set generation and visualization**

| Sequence 5′- 3′ | Purpose | Probe sets used with |
|---|---|---|
| ATTO488-CCAGTGCTCGTGTGAGAAGTC | PCR primer | Xq28 "X.1" |
| CTGCAGAGAAGAGGCAGGTTC | PCR primer | Xq28 "X.1" |
| ATTO565-CGCTCGGTCTCCGTTCGTCTC | PCR primer | Xq28 "X.2" |
| GGGCTAGGTACAGGGTTCAGC | PCR primer | Xq28 "X.2" |
| Alexa647-CAGGTCGAGCCCTGTAGTACGA | PCR primer | Xq28 "X.3" |
| ATTO488-TTGATCTACATATTCAGGTCGAGCCCTGTAGTACG | PCR primer | Xq28 "X.3" |
| CTAGGAGACAGCCTCGGACAC | PCR primer | Xq28 "X.3" |
| TATGTAGATC-CY3B | DNA-PAINT | Xq28 "X.3" |
| TTATACATCTAG | DNA-PAINT | Xist RNA |
| CTAGATGTAT-CY3B | DNA-PAINT | Xist RNA |
| CCATGGCGAGAAGGTCTGCG | PCR primer | "19.1", "19.2", chr19 20 kb |
| TAATACGACTCACTATAGGGGCGTGTGCGAGTGGTTGGAC | PCR primer + T7 | "19.1", "19.2", chr19 20 kb |
| GTGTACCGCGATCCGAAGCGGGTCTTACAGCGGCGCAATGTTCACACGCTCTCCGTCTTGGCCGTGGTCGATCA | Secondary oligo | 19p13.2 "19.1" |
| Alexa647-TGATCGACCACGGCCAAGACGGAGAGCGTGTGAGATGTTT-Alexa647 | Tertiary oligo | 19p13.2 "19.1" |
| GGTGTGGGCTAGCGCCAATCGGTCTTACAGCGGCGCAATGTTTAGCGCAGGAGGTCCACGACGTGCAAGGGTGT | Secondary oligo | 19p13.2 "19.2" |
| CY3B-ACACCCTTGCACGTCGTGGACCTCCTGCGCTA-CY3B | Tertiary oligo | 19p13.2 "19.2" |
| Alexa647-ACACCCTTGCACGTCGTGGACCTCCTGCGCTATTTTTTTT-Alexa647 | Tertiary oligo | 19p13.2 "19.2" |
| CTCCGGCGGACTCATCCCAGGGTCTTACAGCGGCGCAATGTTCACCGACGTCGCATAGAACGGAAGAGCGTGTG | Secondary oligo | 19p13.2 20 kb |
| Alexa647-CACACGCTCTTCCGTTCTATGCGACGTCGGTGTTTTTTTT-Alexa647 | Tertiary oligo | 19p13.2 20 kb |
| Alexa405-CATTGCGCCGCTGTAAGACC | Tertiary oligo | 19p13.2 "19.2" & 20 kb |

**Dataset S1. OligoMiner readme file**

Dataset S1

**Dataset S2. Temperature-dependent differences in off-target discrimination potential**

Dataset S2