Supplementary Methods

Yangho Chen, Tade Souaiaia, and Ting Chen*

1 Multiple Spaced Seed Design

Despite much research [1] [2] [3] [4] [5] [6] [7] has been devoted to the optimization of multiple spaced seeds for different sensitivity criteria, we proposed the following three methods to generate full sensitive periodic multiple seeds. For large genome re-sequencing application, multiple index tables can be queried with the MapReduce framework as proposed in [8] to increase the mapping efficiency and sensitivity by utilizing the higher weight of multiple seeds.

1.1 Design paired periodic seeds with exhaustive search

The design of single periodic seeds can be generalized to find same-length periodic multiple seeds. Tables 1 and 2 show the increase in weight for different period lengths which results from the using paired rather than single seeds. Fig 1 displays the local maximum of weight-length ratios for paired seed periods.

Table 1. The maximum period weight for single and paired seeds at different sensitivity levels

Full Sensitive to	2 substitutions								3 substitutions								4 substitutions							
Repeat length	6	7	8	9	10	11	12	13	6	7	8	9	10	11	12	13	6	7	8	9	10	11	12	13
Repeat weight	3	4	4	5	6	7	8	9	2	2	3	3	4	5	5	6	1	1	1	2	3	3	3	4
Paired repeat weight	3	4	5	6	7	8	8	10	2	3	3	4	5	6	6	7	1	2	2	3	3	3	4	5

Table 2. The maximum weight for single and paired seeds period at SOliD specific sensitivity levels

Full Sensitive to	1 base + 1 color substitutions								2 base substitutions							
Period length	6	7	8	9	10	11	12	13	6	7	8	9	10	11	12	13
Period weight	2	2	3	4	5	5	6	7	1	2	2	3	4	5	5	6
Paired Period weight	2	3	4	5	6	6	7	8	2	3	3	4	5	6	6	7

1.2 IP reduction for finding the weight-maximized paired seeds

We attempted to reduce the problem of maximizing paired spaced seeds to an integer programming problem to generate solutions using lp_solve [9]. For ease of explanation, we give the IP reduction to maximize the weight W of a single seed given a fixed read length $|R|$ and seed length $|\Pi|$.

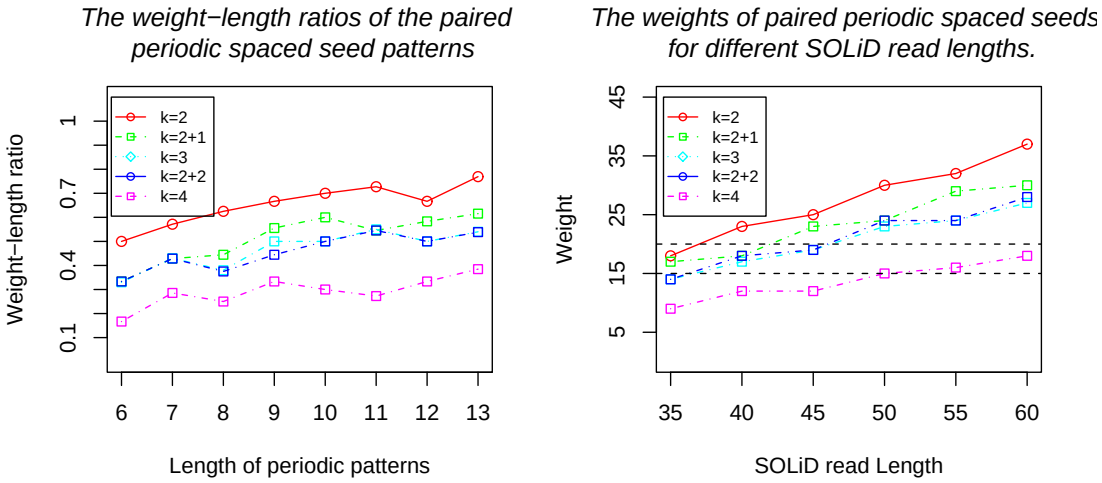

Fig. 1. The left figure displays maximum weight-length ratios for paired seed periods of differing length. The right figure displays the seed weight of periodic paired seeds for different read lengths and sensitivity requirements. We include dashed lines to indicate the minimum and sufficient seeds weights for genome scale mapping, 14 and 20.

Table 3. Paired Seeds with different full sensitivity threshold on 34 color signals

Seed Patterns, parenthesized according to the periodic repeat	Full Sensitivity to mismatch types	Index Per Base	Queries Per Read	Original Weight	Extended Weight	E(Random Hits) Per Reads (3B)
(11111∗∗∗)(11111∗∗∗)(11111∗∗∗)(111) (111∗11∗∗)(111∗11∗∗)(111∗11∗∗)(111)	2 color mis	2	16	18	18 to 22	0.347
(111111∗∗∗∗)(111111∗∗∗∗)(11111) (1111∗∗11∗∗)(1111∗∗11∗∗)(1111∗)	1 base + 1 color mis	2	20	16, 17	16 to 21	4.249
(111∗11∗1∗∗∗)(111∗11∗1∗∗∗)(11) (111∗∗∗1∗11∗)(111∗∗∗1∗11∗)(11)	3 color mis	2	22	14	14 to 19	84.03
(111111∗∗∗∗∗)(111111∗∗∗∗∗)(11) (1111∗∗11∗∗∗)(1111∗∗11∗∗∗)(11)	2 base mis	2	22	14	14 to 19	63.06
(111∗∗∗∗∗∗)(111∗∗∗∗∗∗)(111∗∗∗∗∗) (11∗1∗∗∗∗∗)(11∗1∗∗∗∗∗)(11∗1∗∗∗∗)	4 color mis	2	18	9	9 to 12	110507

.

Maximize W
Subject to

$\sum_{i=1}^{|R|} u_i = W$

for $1 \leq i \leq j \leq |R|$ **and** $l = 0, ..., min\{|R| - |\Pi|, i - 1\}$

$u_{i-l} + x_{ijl} \leq 1$

$u_{j-l} + x_{ijl} \leq 1$

$x_{ijl} \in \{0, 1\}$

for $1 \leq i \leq j \leq |R|$ $\quad \sum_{l=0}^{|R|-|\Pi|} x_{ijl} \geq 1$

for i $= |\Pi| + 1, ..., |\mathrm{R}|$ $\quad u_i = 0$

for i $= 1, ..., |\Pi|$ $\quad u_i \in \{0, 1\}$

$u_i = 1$ if position i on the seed is a "care position", otherwise $u_i = 0$.

$\forall i, j$ the spaced seeds can shift $|R| - |\Pi|$ times with l representing a specific shift, s.t. at least one flag x_{ijl} is covered. The weight is equal to the sum of u_i.

Providing further justification for the use of periodic seeds, lp_solve returned the same single spaced seeds, F_2 and F_3, that we designed. In theory this reduction can be easily generalized to

many seeds for a target function which maximizes the lowest the seed weights. Unfortunately, the IP reduction is too slow in practice to find solutions containing three or more seeds.

1.3 Multiple periodic seed design method: mismatch tuple grouping

In addition to the methods described above we propose a simple algorithm to design full sensitive multiple seeds by grouping all possible mismatches tuples. Our simple strategy demonstrates a generalizable seed design scheme, which often provides better results than methods that divide reads into equal size fragments.

For example, to design multiple seeds full sensitive to three mismatches with period length as 5, all the possible patterns, $\binom{5}{3}$ are list as follows. These patterns are assigned into two groups corresponding to two seeds. Each pattern in a group can be transformed to the other pattern by performing shifting the pattern along the length of the read.

1) (11***), (1***1), (***11), (**11*), (*11**)
2) (1*1**), (*1**1), (1**1*), (**1*1), (*1*1*)

For read with length 32, the full sensitive seed pair is

(11***)(11***)(11***)(11***)(11***)11*
(1*1**)(1*1**)(1*1**)(1*1**)(1*1**)1*1

It can also be extended to utilize the full read length as the single seed.

(11***)(11***)(11***)(11***)(11***)(11***)11
(1*1**)(1*1**)(1*1**)(1*1**)(1*1**)(1*1**)1*

Ten queries is needed. Each query has the effective seed weight from 12 to 14. This is a better choice than the full sensitive multiple seed in C_3 family, which also needs ten queries with weight 12, however with four instead of two index tables.

(111111)(111111)
(111111)(******)(111111)
(111111)(******)(******)(111111)
(111111)(******)(******)(******)(111111)

Let $M(p,k)$ denote the multiple seed designed by grouping $\binom{p}{k}$ mismatches tuples, where p is the period length and k is the mismatches number of full sensitivity threshold. Similarly, let $MS_{1,1}$ be the special seed family full sensitive to one color and one base used for the SOLiD reads. The following table compare the M and MS seed families with the single periodic seed family F and C family which divide reads into fragments.

The expected random hits number per read is, the number in each cell multiplied by the reference length that the read will be mapped to. This table indicates the practical limits, in terms of number of tables, queries and expected random hits of using multiple index tables to achieve full sensitivity to four and five mismatches.

2 Analysis of Periodic Seeds for Weight Optimization

The high seed weight achieved through the use of periodic space seeds agrees with the results from Kucherov [7], that weight optimization usually involves a periodic seed. We design our seeds through weight optimization of a repeating pattern, a method that yields high weight, generalizable and extendable spaced seeds. However, it does not preclude that there exists a non-periodic single spaced seed of higher weight. Here we show that F_2 generalized to a 34bp achieves the maximum possible seed weight which provides full sensitivity to two mismatches in seven slides.

Table 4. Comparison of single and multiple seeds generated by the mismatch tuple grouping method

Seed family	# Index tables	# Queries	Expected Random Hits Per Read / Reference Length				
			$\|R\| = 25$	$\|R\| = 30$	$\|R\| = 32$	$\|R\| = 34$	$\|R\| = 36$
F_2	1	7	1.99e-007	2.63e-009	7.79e-010	3.15e-010	2.48e-011
M(5, 2)	2	10	4.56e-008	7.13e-010	1.40e-010	3.98e-011	3.98e-012
M(6, 2)	3	15	2.04e-008	2.54e-010	3.63e-011	6.43e-012	9.91e-013
M(7, 2)	3	21	1.49e-008	8.00e-011	1.46e-011	3.66e-012	2.48e-013
M(8, 2)	4	28	2.12e-008	1.36e-010	1.97e-011	1.53e-012	2.21e-013
C_2	3	6	3.58e-007	2.24e-008	1.40e-009	1.40e-009	8.73e-011
$S_{1,1}$	1	10	7.95e-006	3.28e-007	3.01e-008	1.15e-008	4.04e-009
$MS_{1,1}(6,3)$	3	18	2.01e-006	7.54e-008	2.04e-008	6.68e-009	1.18e-009
$MS_{1,1}(7,3)$	4	28	9.78e-007	1.33e-008	3.82e-009	1.27e-009	1.10e-010
$MS_{1,1}(8,3)$	5	40	4.61e-007	1.09e-008	1.61e-009	1.60e-010	4.70e-011
$MS_{1,1}(9,3)$	6	54	8.36e-007	3.83e-009	8.59e-010	2.04e-010	2.96e-011
F_3	1	11	3.08e-005	3.20e-006	1.30e-006	1.05e-007	3.01e-008
M(5, 3)	2	10	1.81e-005	1.13e-006	4.06e-007	1.83e-007	4.28e-008
M(6, 3)	4	20	3.58e-006	1.15e-007	3.02e-008	9.12e-009	1.79e-009
M(7, 3)	5	35	1.36e-006	1.95e-008	5.31e-009	1.60e-009	1.48e-010
M(8, 3)	7	56	7.04e-007	1.53e-008	2.32e-009	2.80e-010	6.59e-011
C_3	4	10	9.54e-006	5.96e-007	5.96e-007	5.96e-007	3.73e-008
F_4	1	10	1.34e-003	1.41e-004	6.60e-005	3.60e-005	1.72e-005
M(5, 4)	1	5	4.88e-003	1.22e-003	8.54e-004	4.88e-004	2.59e-004
M(6, 4)	3	15	3.78e-004	3.43e-005	1.56e-005	7.51e-006	2.15e-006
M(7, 4)	5	35	7.72e-005	3.02e-006	1.21e-006	4.43e-007	7.66e-008
M(8, 4)	10	70	2.79e-005	1.44e-006	2.58e-007	6.08e-008	1.79e-008
C_4	5	15	2.29e-004	1.43e-005	1.43e-005	1.43e-005	8.94e-007
M(6, 5)	1	6	2.05e-002	5.86e-003	4.39e-003	2.93e-003	1.46e-003
M(7, 5)	3	21	2.66e-003	2.95e-004	1.66e-004	8.01e-005	2.65e-005
M(8, 5)	7	56	5.37e-004	6.75e-005	1.50e-005	5.34e-006	2.58e-006
M(9, 5)	14	126	5.21e-004	1.60e-005	5.83e-006	2.03e-006	3.72e-007
C_5	6	21	5.13e-003	3.20e-004	3.20e-004	3.20e-004	2.00e-005

Definition 1. *Let M_2 represent that maximum weight fixed length single seed to provide full sensitivity to two mismatches after seven slides for a 34bp read. M_2 is some combination of twenty-eight "care" ("1") and "don't care" positions ("*").*

For M_2 to provide full sensitivity it must cover twenty-eight positions pairwise with "*" during the seven slides. Thus, the following inequality must hold.

$\binom{28}{2} \leq \binom{28-w}{2} * 7$

This inequality holds only for values of w below eighteen, meaning M_2 cannot have weight greater than seventeen. However, if we consider that adjacent positions must also be covered pairwise with ("*"), we can observe the natural inefficiency which results from allowing just seven slides. The consecutive "*" positions required every seven positions to cover all adjacent pairs will cover $\binom{4}{2}$ redundant positions each of the six times the seed slides. Subtracting these thirty-six redundant pairs to the above inequality results in w no greater than sixteen. Thus, M_2 can have weight no greater than sixteen, equal to the weight of F_2.

3 Estimation of partial sensitivity

The partial sensitivity is calculated by mapping simulated reads. Ten million reads are simulated from the human chromosome one with exactly four and five mismatches. 88.1% and 66.8% of

the reads can be mapped back to chromosome one, within four and five mismatches threshold respectively using the F_3 seed. 91.1% and 69.7% of these simulated reads can be mapped back to the whole human genome, with the F_3 seeds.

References

[1] François Nicolas and Eric Rivals, "Hardness of optimal spaced seed design", *J. Comput. Syst. Sci.*, vol. 74, no. 5, pp. 831–849, 2008.

[2] M Li, B Ma, D Kisman, and J Tromp, "Patternhunter II: highly sensitive and fast homology search", *J Bioinform Comput Biol*, vol. 2, no. 3, pp. 417–439, Sep 2004.

[3] "Efficient methods for generating optimal single and multiple spaced seeds", in *BIBE '04: Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*, Washington, DC, USA, 2004, p. 411, IEEE Computer Society.

[4] L Noé and G Kucherov, "Improved hit criteria for dna local alignment", *BMC Bioinformatics*, vol. 5, pp. 149–149, Oct 2004.

[5] Y Sun and J Buhler, "Designing multiple simultaneous seeds for dna similarity search", *J Comput Biol*, vol. 12, no. 6, pp. 847–861, Jul-Aug 2005.

[6] H Lin, Z Zhang, M Q Zhang, B Ma, and M Li, "Zoom! zillions of oligos mapped", *Bioinformatics*, vol. 24, no. 21, pp. 2431–2437, Nov 2008.

[7] G Kucherov, L Noé, and M Roytberg, "Multiseed lossless filtration", *IEEE/ACM Trans Comput Biol Bioinform*, vol. 2, no. 1, pp. 51–61, Jan-Mar 2005.

[8] M C Schatz, "Cloudburst: highly sensitive read mapping with mapreduce", *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, Jun 2009.

[9] Michel, Jeroen, Kjell, and Peter, "Lp_solve", 2007.