# OptimusPrime: Multiplex Primer Design Tool for Hi-Plex Sequencing

*Edmund Lau[1], Daniel Park[1,2] , Bernard Pope[1,3,*]*

www.hiplex.org

## Introduction

The advent of Massively Parallel Sequencing (MPS) has dramatically reduced the cost and increased the throughput of DNA sequencing. A number of methods have been developed that target specified genomic regions for MPS. However, these, variously, are compromised by issues of relative expense, accuracy, requirement for specialist equipment and the cumbersome nature of protocols. In this work we present Hi-Plex, a novel MPS platform, and multiplex Polymerase Chain Reaction (PCR) primer design software to meet the challenges the platform presents.

1.

## What is Hi-Plex?

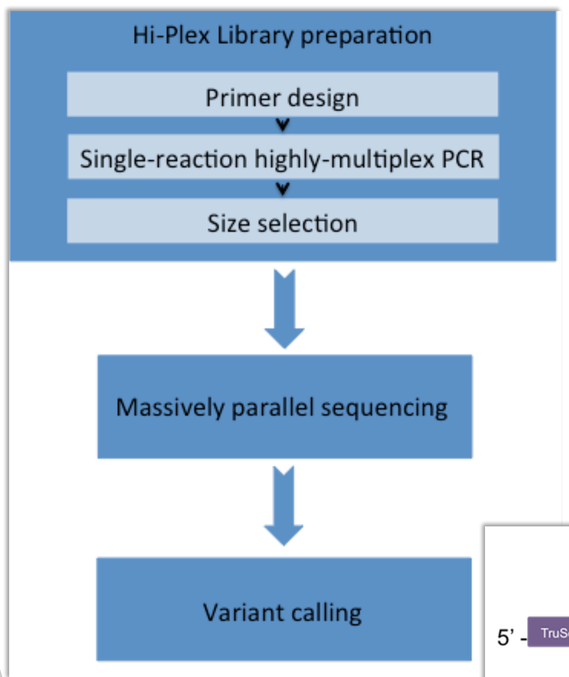
Hi-Plex is a Massively Parallel Sequencing (MPS) platform using a highly multiplexed PCR-based approach for targeted MPS that employs tagged gene-specific-primers (GSP) to seed and universal bridge primers to drive the majority of amplification, minimising bias caused by differing gene-specific primer efficiencies. Compared to traditional sequencing methods, Hi-Plex has the following advantages:

***Simple operating protocol***

Library preparation for Hi-Plex only requires a single multiplex PCR using equipment readily available in standard labs. Hi-Plex requires no automation, specialist machinery or expensive reagents for streamlined high-throughput screening.

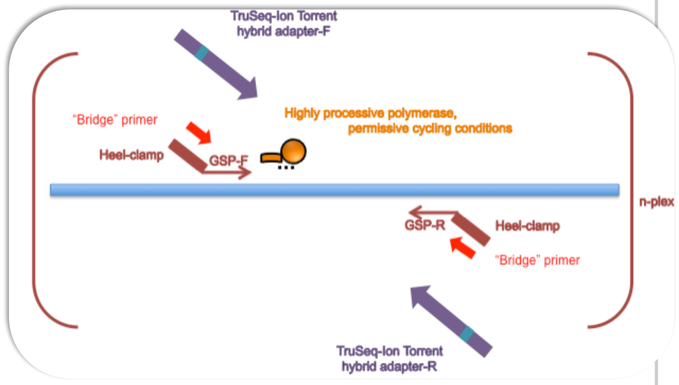
***Cost-effective reagents***

No specialist reagents – simply unmodified oligonucleotides and DNA polymerase yielding a highly cost-effective system.



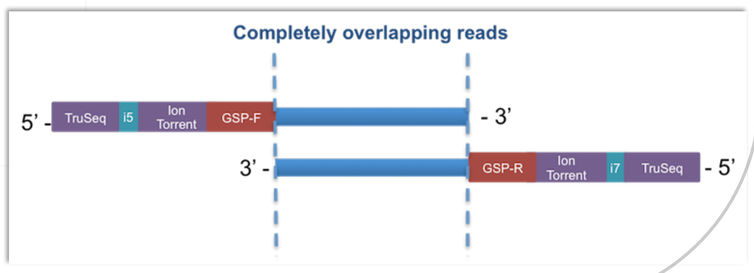

***Supports the specification of small and tight-ranged insert size***

The size selection step allows the target insert size to be defined within a small and tight range. This increases uniformity of amplification efficiency, provides a filter for off-target artefacts and opens the platform to the use of degraded DNAs, such as those derived from FFPE-tumour specimens.




***Highly Accurate***

Hi-Plex's PCR reaction is compatible with the highest fidelity thermostable DNA polymerases available. A small insert size allows the read-pair to overlap completely allowing the application of read-pair-overlap-considerate variant calling software (e.g. UNDR-ROVER) to perform variant calling with high statistical confidence, sensitivity and accuracy.

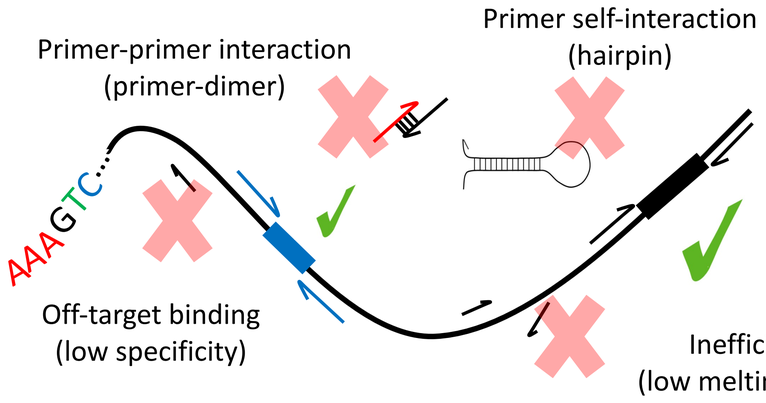



2.

## Challenges in Multiplex-PCR

While Hi-Plex is tolerant to reagent design defects that would preclude the use of conventional multiplex-PCR, it is not immune to very strong inhibitory effects.

***What can go wrong in multiplex-PCR?***




***What do we expect?***

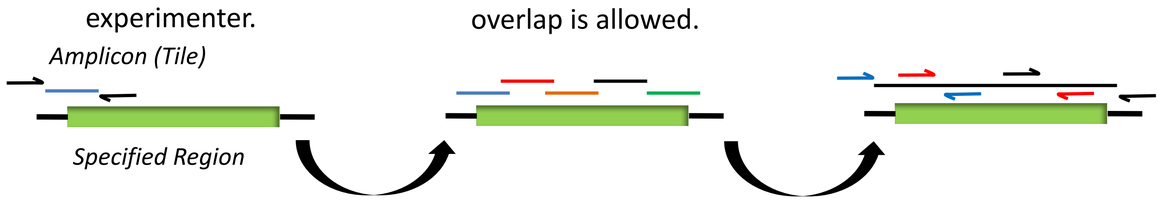
The ideal experimental outcome of multiplex-PCR will be a uniform amplification of all specified regions with no other undesired products.

***Formulation of the multiplex primer design problem***

Every pair of forward and reverse gene specific primers defines an amplicon, or a "tile". Allowed tile sizes are specified by the experimenter.

The solution space consists of all possible tiling patterns. A candidate solution is a tiling that covers the region. A specified amount of tile overlap is allowed.

All candidate solutions are compared. The tiling that maximises the chance of a successful reaction is output as a solution.




***A problem with SIZE – the need for an efficient primer design tool!***

The number of candidate solutions easily exceeds the millions and *grows exponentially* with region size. This means a brute force computational search will not be able to find optimal solution in reasonable time.

3.

## A Computational Solution

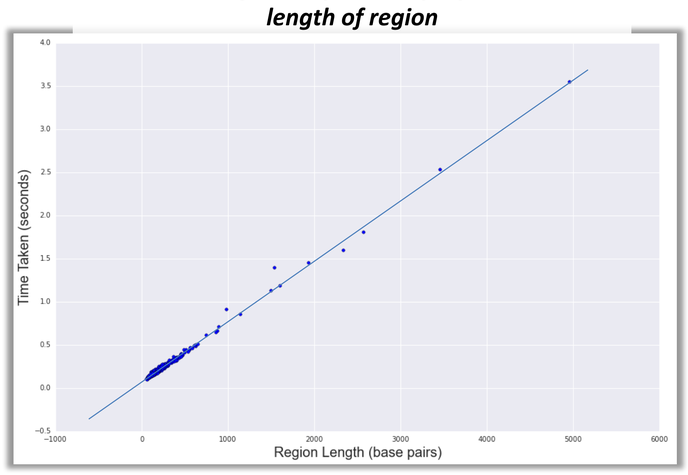
The primer design problem is framed as a computational optimisation problem. We employ a search-and-score technique to explore the solution space and output the best solution.

### Search

***Dynamic Programming (DP)***

By reformulating the problem as a recursive multistage decision problem, we are able to use dynamic programming to explore the exponentially large search space in linear time. With this formulation, the output solution is provably optimal with respect to a given tile-based scoring.

***Time taken for computation proportional to length of region***



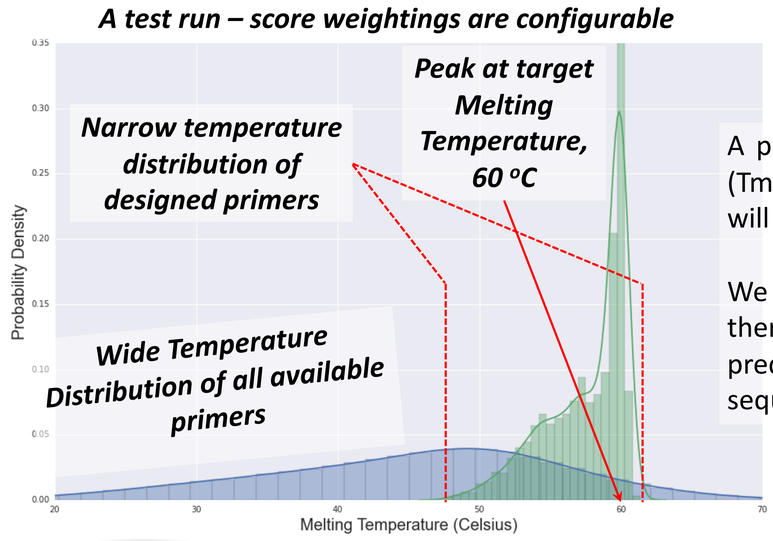

***A recursive problem***

"The best choice for the next tile is the choice that maximises the score given the choices made so far."

### Score

While exploring the solution space, the algorithm will be presented with individual primers and asked to evaluate its suitability as part of the final solution. This is done by assigning several different scores to criteria relevant to the success of a reaction. These scores are then combined via a sum with user configured weightings to obtained a goodness measure for any primer candidate.

***A test run – score weightings are configurable***




***Melting Temperature***

A primer with Melting Temperature (Tm) close to the specified target Tm will score higher.

We use the Nearest-Neighbour DNA thermodynamic model to accurately predict Tm given a primer sequence.

***Specificity***

The Burrow Wheeler Alignment (BWA) algorithm is employed to predict potential binding sites for a given primer sequence.
This allows elimination of primers with the potential to bind off-target.

***Secondary Structure***

***Runs*** ***Hairpins*** ***Entropy*** ***%GC***

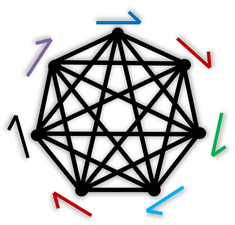
Several different measures are implemented to avoid a chosen primer forming undesirable secondary structures.
These scores ensure sufficient complexity of the chosen sequence and avoid homologous subsequences.

### Post-processing & Output

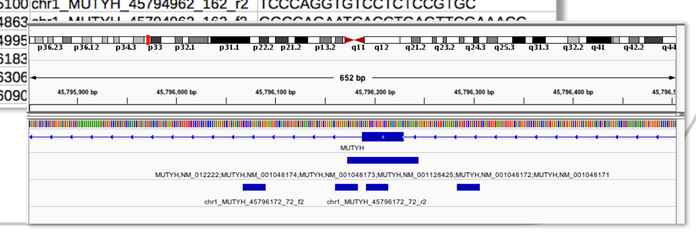
***Primer Dimer: A Hard Problem***

Detection of potential primer-dimer conflicts requires all-to-all comparison of primers in the candidate solution. Incorporating this computation into the optimisation is known to be a computationally hard problem. Thus, we post-process a given solution, looking for potential conflicts and redesign the offending primers. The final solution is output in tabular form and is readily processed and visualised by standard bioinformatics tools.

***All-to-all comparison***



| chrom | start | end | name | sequence |
|---|---|---|---|---|
| chr1 | 45795049 | 45795068 | chr1_MUTYH_45794962_162_f1 | CTGGCCCATGCGGGGCTTT |
| chr1 | 45795170 | 45795189 | chr1_MUTYH_45794962_162_r1 | CGGGGAAAGGGAGAGAGGACAAGG |
| chr1 | 45794942 | 45794966 | chr1_MUTYH_45794962_162_f2 | ACAGGATTCTCAGGGAATGGGGGC |
| chr1 | 45795076 | 45795100 | chr1_MUTYH_45794962_162_r2 | TCCCAGGTGTCCTCTCCGTGC |
| chr1 | 45794837 | 45794863 | | |
| chr1 | 45794969 | 45794995 | | |
| chr1 | 45796160 | 45796183 | | |
| chr1 | 45796283 | 45796306 | | |
| chr1 | 45796067 | 45796090 | | |




***Acknowledgement***

This project was supported by VLSCI and The Department of Pathology at The University of Melbourne as part of the Undergraduate Research Opportunity Program (UROP) Biomedical Research Victoria. Further support was provided via Cancer Council Victoria grant-in-aid #1066612 and NHMRC project grant #1108179.

***References***

Nguyen-Dumont, T. et al. Anal. Biochem. 442, 127–129 (2013).
Nguyen-Dumont, T. et al. Biotechniques 55, 69–74 (2013).
Nguyen-Dumont, T. et al. Biotechniques 58, 33–36 (2015).
Nguyen-Dumont, T. et al. BMC Med. Genomics 6, 48 (2013).
Nguyen-Dumont, T. et al. Anal. Biochem. 470, 48–51 (2015).
Pope, B.J. et al. Source Code Biol. Med. 9, 3 (2014).
Park, D.J. et al. BMC Bioinformatics 17, 1–7 (2016).

***Affiliations***

1 Victorian Life Science and Computation Initiative, The University of Melbourne
2 Department of Pathology, The University of Melbourne
3 Department of Computing and Information Systems, The University of Melbourne
** Corresponding author (bjpope@unimelb.edu.au)*