

From: [Eric Boutin](#)
To: "Patrick Waller"
Cc: greg.porreca@gmail.com; ulaserson@gsgenetics.com; "Bob Wassman"; "Paris Wallace"; jim.billy.ii@gmail.com
Subject: Good Start Genetics - IP Meeting Wed 11th at 9am
Date: Friday, March 6, 2009 3:11:09 PM
Attachments: [Bias-detecting_Barcodes_v2.pdf](#)
[gsg_patapp_reduce-amp-bias2.pdf](#)
[ProvIP_MIP_length_20090307.pdf](#)
[bias_detecting_barcodes_fig1.pdf](#)
[bias_detecting_barcodes_fig2.pdf](#)

Patrick,

Our team is looking forward to meeting with you next Wednesday and talking in greater detail about our provisional patent applications, and about our royalty/ licensing strategy around certain disorders.

Below is an email from Greg Porreca, our Director of Technology, outlining thoughts/ questions we'd like to consider. Please feel free to reach out to Greg directly if you have any questions in advance of the meeting.

Thank you,
Eric

Eric Boutin
Good Start Genetics | Founder
Mobile (646) 382-4421
eboutin@gsgenetics.com

From: Gregory Porreca [<mailto:greg.porreca@gmail.com>]
Sent: Friday, March 06, 2009 2:37 PM

Hello Patrick-

Please find drafts of our three provisional patent applications attached. Below is also a list of papers and patents associated with the drafts.

Aside from next steps working with you to refine the document and add information where needed, we'd like to discuss several topics on Wednesday:

- 1) have we defined the inventions broadly enough (or, is it possible to increase breadth further)?
- 2) how much overlap do you see in the "Bias-detecting barcodes" application with the patent reference called out in Note 1 on page 1 (McCloskey et al.)
- 3) how clear are the figures?

McCloskey et al., US Patent Application Publication number: US 2007/0020640 A1

Miner et al. 2004 Nucl. Acids Res. 32:e135

McCloskey et al. 2007 Biochemical Genetics 45:1-6

Olson 2007 Nature Meth 4:891

Hardenbol et al. 2005 Genome Res 15:269-75

Porreca et al. 2007 Nat Methods 4:931-6

Krishnakumar et al. 2008 Proc Natl Acad Sci USA 105:9296-301

Shendure & Ji 2008 Nature Biotech 26:1135-45

--



ML-IPR-000001

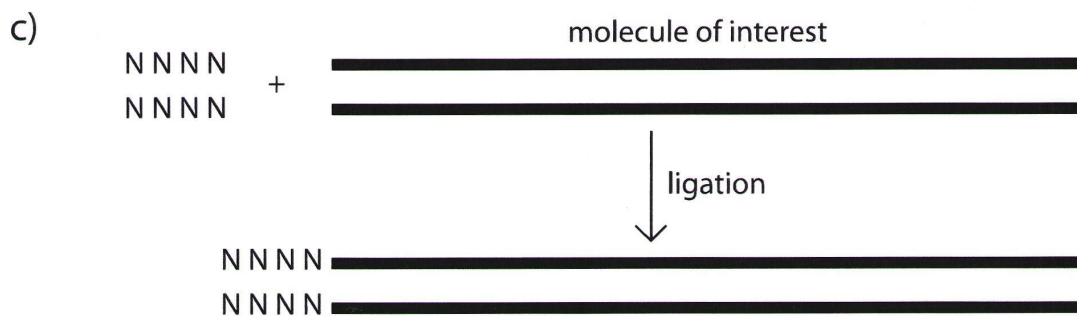
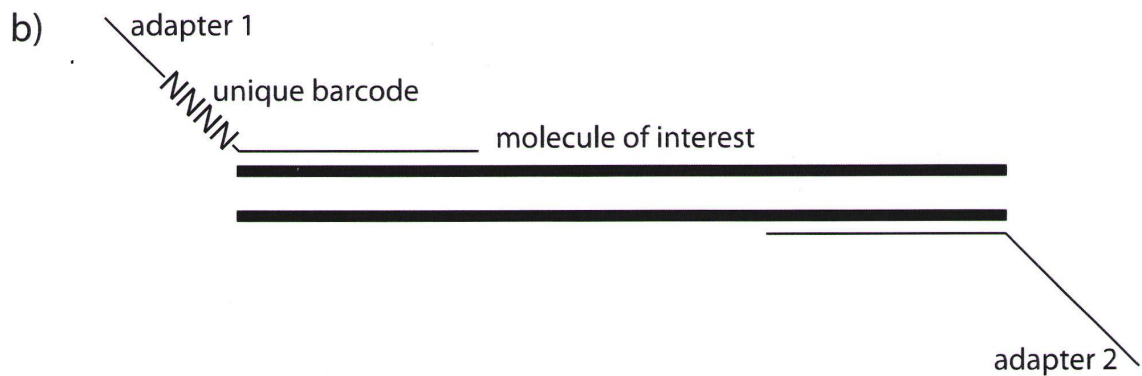
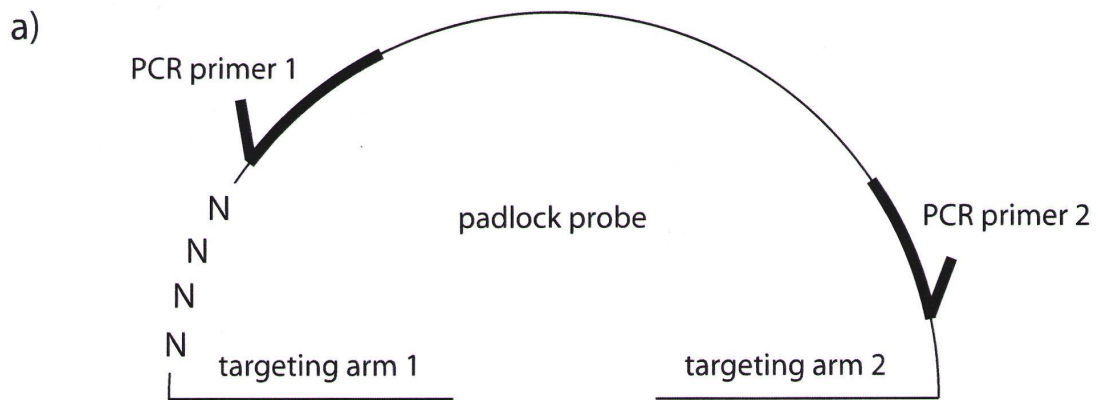
Gregory Porreca, Ph.D.
Lecturer on Genetics

Church Laboratory
Dept. of Genetics, Harvard Medical School
New Research Building 260
77 Avenue Louis Pasteur
Boston, MA 02115

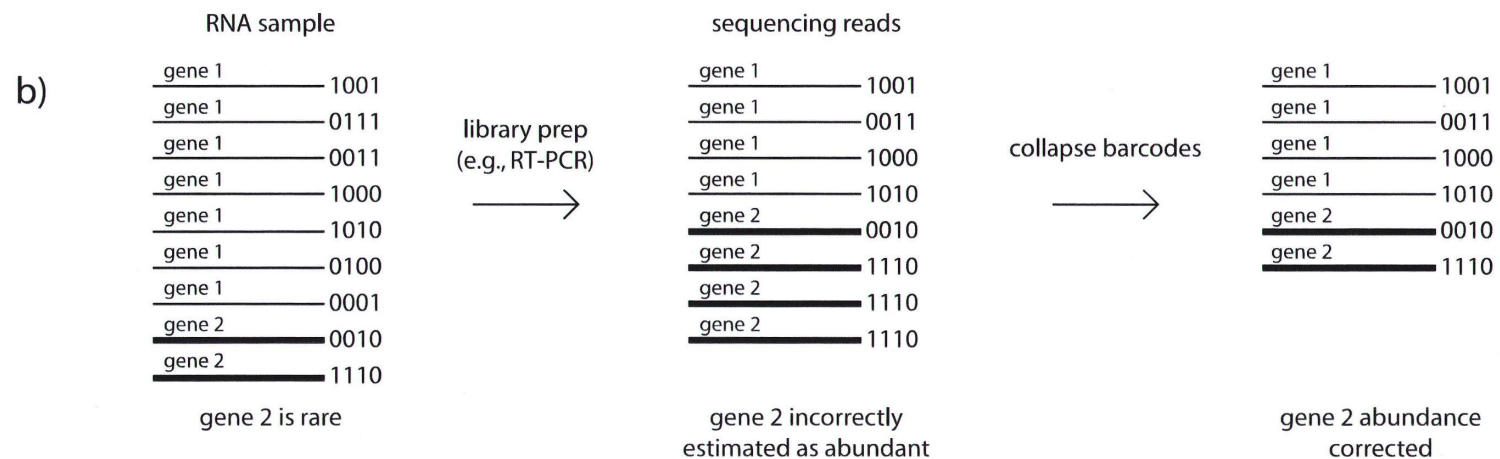
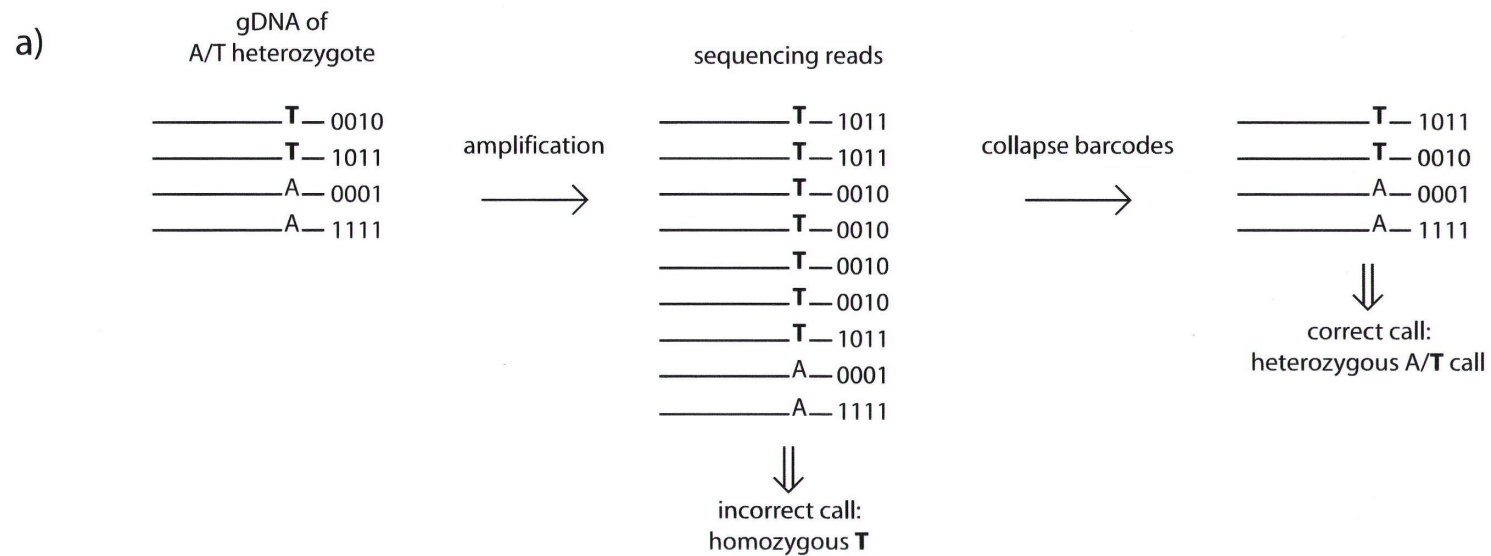
tel: (617) 432-7172
fax: (617) 432-6513

cv: http://arep.med.harvard.edu/CVs/Porreca_CV06.pdf

ML-IPR-000002



ML-IPR-000003



ML-IPR-000004

METHODS FOR REDUCING BIAS IN PREPARATIONS OF NUCLEIC ACIDS

Field of the Invention

The invention relates to preparative steps in DNA sequencing-related technologies that reduce bias and increase the reliability and accuracy of downstream quantitative applications.

Background of the Invention

There are currently many genomics assays that utilize next-generation (polony-based) sequencing to generate data, including genome resequencing, RNA-seq for gene expression, bisulphite sequencing for methylation, and Immune-seq, among others. In order to make quantitative measurements (including genotype calling and gene expression profiling), these methods utilize the counts of sequencing reads of a given genomic locus as a proxy for the representation of that sequence in the original sample of nucleic acids.

All of these techniques require a preparative step to construct a high-complexity library of DNA molecules that is representative of the sample of interest. This may include chemical or biochemical treatment of the DNA (e.g., reverse transcription, bisulphite treatment), capture of a specific subset of the genome (e.g., padlock probe capture, solution hybridization), and a variety of amplification techniques (e.g., polymerase chain reaction, whole genome amplification, rolling circle amplification).

A common problem to overcome is that these techniques commonly introduce biases into the sequencing libraries, generally due to biases in enzyme function or to the exponential and stochastic nature of amplification methods like PCR. As most of the downstream quantification techniques depend on stochastic counting processes, these problems are typically overcome through a brute-force approach: sampling enough to obtain the minimum number of observations necessary to make statistically significant decisions. This strategy of oversampling is generally limited to elimination of low-count Poisson noise. At best, this approach wastes resources and increases the expense required to perform an experiment; at worst, it reduces the statistical confidence in the conclusions that are reached based on the data.

As part of the library preparation process, adapters are commonly ligated onto the ends of the molecules of interest. The adapters

ML-IPR-000005

typically contain any of PCR primer sites (for amplification or emulsion PCR), sequencing primer sites, and molecular barcodes that uniquely identify individual samples that may be mixed together. Barcodes with defined sequences have been used in multiple contexts to simplify experiments or increase the cost-effectiveness. Common examples include multiplexing multiple samples together in a single reaction and barcoding features that need to be identified later (e.g., chromosome conformation capture) by sequencing short defined segments.

[NOTE1: There is some prior art associated with random barcodes, but I'm not sure if it anticipates the applications we are interested in. See, mainly, USPTO Publication Number US 2007/0020640 A1 (McCloskey et al.).]

[NOTE2: We can put a lot more detail here regarding the specific biasing problems described, such as how this will affect genotype calling, RNA expression level quantitation, etc., and exactly why biasing causes a problem here.]

Summary of the Invention/Detailed Description [We can break it up later.]

The invention is the composition and use of a large library of unique molecular barcodes in preparative steps for reducing the effects of the bias described above. During the library preparation process, each molecule is tagged with a unique in such a configuration such that the barcode is sequenced along with the information of interest for which the library is being prepared. The incorporation of the barcode at a given step allows the detection and correction of biases in any of the following steps of the protocol.

One possible method to create a large library of unique barcodes is to use degenerate, random-sequence barcodes of defined length. The sequences of the barcodes will be read at the final stage of the sequencing. The observations of the random barcode sequences are used to detect and correct biases in the final read-out of the library. The total possible number of such random barcodes is 4^N , where N is the length of the barcode. The length of the barcode is adjusted such that the size of the population of unique barcodes causes any given molecule-barcode combination to be very rare; since the barcodes are randomly incorporated, the probability that two individual molecules receive the same barcode sequence is very small. (In fact, the likelihood that the same molecule barcode-combination occurs more than once is even smaller.)

ML-IPR-000006

In the final readout of the sequencing process, each read has an additional unique barcode sequence. Since the barcodes were distributed randomly, we should expect that all the unique barcode sequences should be observed about an equal number of times. (In fact, the number of counts of a given barcode should have a Poisson distribution [Right? Assuming we had a uniform distribution of barcodes that were amplified, what happens to their distribution post-PCR if the efficiencies are unequal? I suppose this also depends on the distribution of the efficiencies.]). If any barcodes or barcode-molecule combinations are statistically overrepresented, it is a sign that there was some kind of bias in the protocol between the incorporation of the random barcodes and the actual sequencing.

The number of reads of a given sequence is typically taken as a proxy for the amount of that sequence present in the originating sample. The distribution of sequences in the originating sample is the quantity of interest. However, as the preparative steps taken during the protocol are known to cause biases and changes in the distribution (particularly for relatively rare events), the fidelity of using the count numbers is often questionable. Using the random barcode method, the distribution of the random barcodes is well known at the time of incorporation (it is uniform); by observing how the distribution of random barcodes changes after the sequencing is performed, we have collected the information that is necessary to correct the biases. By undoing the distribution changes that are measured for the random barcodes, the corresponding changes in the counts for the sequences of interest will provide a higher fidelity representation of the composition of the original complex pool. Furthermore, the distribution of the unique library of barcodes may be biased as well. However, the library is designed to be significantly larger than the pool of molecules of interest; this ensures that the incorporation of the barcodes is significantly undersampling the population of unique barcodes, ensuring that distribution of unique barcodes that was actually incorporated is effectively uniform.

For example, a typical experiment utilizing 1 µg of gDNA will contain about ~150,000 copies of a diploid genome. If we incorporate a library of random 12-mer barcodes (~1.6 million barcodes), there would be more than 100 unique sequences per genomic copy. If we incorporate random 15-mers (~1 billion barcodes), there are more than 7,000 unique sequences per genomic copy. Therefore, the probability of the same barcode getting incorporated multiple times is incredibly small.

ML-IPR-000007

The actual incorporation of the random barcodes can be performed through various methods. One method is to incorporate the tags by ligation. This is very flexible as the molecules can be ligated to any blunt-ended nucleic acids. The sequencing primers must be incorporated afterwards such that they sequence both the barcode and the molecule of interest. Alternatively, the sequencing adaptors can be synthesized with the random barcodes at their 3' end (as degenerate bases), so that only one ligation must be performed. Another method is to incorporate the barcode into a PCR primer, such that the primer structure is arranged with the common adaptor sequence followed by the random barcode followed by the PCR priming sequence (in 5' to 3' order). The random barcode and adaptor sequence (which may contain the sequencing primer site) get incorporated as tags. Another method to incorporate the barcodes is to synthesize them into a padlock probe prior to performing a gene capture reaction. The random barcode is incorporated 3' to the targeting arm but 5' to the amplification primer that will be used downstream in the protocol. Another method to incorporate the random barcodes is as a tag on a gene-specific or poly-dT reverse-transcription primer. This allows the barcodes to be incorporated directly at the cDNA level.

At the incorporation step, the distribution of barcodes can be assumed to be uniform. At that point, any biases in any part of the protocol would change the uniformity of this distribution, which can be observed after sequencing. This allows the random barcodes to be used in any preparative process where the ultimate output is sequencing of many molecules in parallel. [Is that the most general statement I can make? Is there something more inclusive we could say instead?]

There may be concern that biasing in the pool of unique barcodes will render this method invalid. However, the incorporation of the barcodes can be performed in a way that minimizes this risk. For example, if the barcodes are incorporated into probes when they are synthesized on-chip in parallel, then the degeneracy of the incorporated nucleotides should be enough to ensure near-uniform distribution. The real concern is when the incorporation occurs as a separate step. The main concern is that the pool of unique barcodes may itself have been amplified, inducing biases in the initial pool. However, in most practical cases, the scale of the column syntheses is large enough so that amplification of this pool is not necessary. By avoiding any type of amplification or selection step on the pool of unique barcodes, we can minimize potential biases.

ML-IPR-000008

One example of the use of the barcodes is in genome resequencing. Considering that the raw accuracy of most next-generation sequencing instruments is relatively low, it is crucial to oversample the genomic loci of interest. Furthermore, since there are two alleles at every locus, it is important to sample enough to ensure that both alleles have been observed a sufficient number of times to confidently determine whether the sample is homozygous or heterozygous. Indeed, the sequencing is performed to sample the composition of molecules in the originating sample. However, after multiple reads have been collected for a given locus, it is possible that due to bias (e.g., caused by PCR amplification steps), a large fraction of the reads derive from a single originating molecule. This would skew the population of genotypes observed, and would affect the outcome of the genotype call. For example, it is possible that a locus that is heterozygous is called as homozygous, because there are only a few observations of the second allele out of many observations of that locus. However, if we also have information on random barcodes, this situation could be averted, because the over-represented allele would be seen to also have an over-represented random barcode (i.e., the sequences with the overrepresented barcode all originated from the same single molecule). Therefore, the sequences and corresponding distribution of random barcodes can be used as an additional input to the genotype-calling algorithm to significantly improve the accuracy and confidence of the genotype calls.

Another embodiment of the invention is in the quantitative analysis of gene expression. When RNA is sequenced, the relative numbers of counts for different transcripts is used as a proxy for the relative abundance of the transcript in the original sample of RNA. As described above, the library preparation protocols may introduce biases that change the relative amounts of the RNA species in the sample that actually gets sequenced (post-processing). However, if each RNA/cDNA molecule is coupled to a unique random barcode sequence, then any repeated barcodes in the sequencing reactions can simply be collapsed into each other. By collapsing similar clones in such a way that the sequencing data only includes a single representative for each barcode (or barcode-molecule combination), the biases that were introduced in preparation are essentially eliminated.

Finally, there is an approach to utilize this random barcode methodology retrospectively on samples that do not explicitly incorporate unique barcodes. Most oligonucleotide synthesis methods are error-prone. The fraction of oligos with errors depends on the length of the

ML-IPR-000009

oligonucleotide and the synthesis method employed. For example, using traditional column synthesis of 25-mers, we would expect that 1 in 40 PCR primers would contain a synthesis error (assuming one error every 1,000 bases). These error-containing primers are also incorporated in the PCR reaction. Since the errors are distributed in a random fashion, they are unique and rare; therefore, the error-containing PCR primers can be considered as unique random barcodes as well. Since these primers typically get sequenced as well, the errors can be detected and used as random barcodes to look for biases in the same way described above.

ML-IPR-000010

Method for reducing representational bias in multiplex amplification of nucleic acids

Field of the invention

The invention relates to methods that reduce bias and increase reproducibility in multiplex amplification of nucleic acids by 'molecular inversion probe' methods.

Background of invention

Molecular inversion probe technology is used to detect or amplify particular nucleic acid sequences in potentially complex mixtures. However, amplification efficiency is non-uniform and not highly reproducible, limiting the preparative utility of the method in research and diagnostic settings.

Use of molecular inversion probes has been demonstrated for detection of single nucleotide polymorphisms (Hardenbol et al. 2005 Genome Res 15:269-75) and for preparative amplification of large sets of exons (Porreca et al. 2007 Nat Methods 4:931-6, Krishnakumar et al. 2008 Proc Natl Acad Sci USA 105:9296-301). In both cases, oligonucleotide probes are designed which have ends ('targeting arms') that hybridize up-stream and down-stream of the locus that is to be amplified. Following hybridization, polymerase fill-in and ligation reactions are performed to convert the hybridized probe to a covalently-closed, circular molecule containing the desired target. PCR or rolling circle amplification plus exonuclease digestion of non-circularized material is performed to isolate and amplify the circular targets from the starting nucleic acid pool. Since one of the main benefits of the method is the potential for a high degree of multiplexing, generally thousands of targets are captured in a single reaction containing thousands of probes.

Summary / detailed description

The invention consists of 1) methods to increase the uniformity of amplification efficiency when multiple targets are amplified in parallel; 2) methods to increase the reproducibility of amplification efficiency; 3) methods to reduce the contribution of targeting oligonucleotide variability to amplification efficiency; 4) methods to reduce the effect on a given target of polymorphisms in probe hybridization regions; 5) methods to simplify downstream workflows when multiplex amplification by MIPs is used as a preparative step for analysis by nucleic acid sequencing.

Current methods design a single probe for each target (a target can be as small as a single base or as large as a kilobase or more of contiguous sequence). While probes are generally designed to meet certain constraints (e.g. melting temperature, G/C content, etc) known to partially affect capture/amplification efficiency, no known set of constraints is sufficient to guarantee either largely uniform or highly reproducible capture/amplification efficiency. Uniformity and reproducibility can be increased by designing multiple probes per target, such that each base in the target is captured by more than one probe. We design multiple MIPs per target to be captured, where each MIP in a set designed for a given target has 'targeting arms' which hybridize to (at least partially) different sequence in (or immediately flanking) the target. In this way,

ML-IPR-000011

differences in efficiency between different targeting arms and fill-in sequences are averaged across multiple MIPs for a single target, which results in more uniform and reproducible capture efficiency.

Polymorphisms in the target sample under the regions flanking the target (i.e., the regions to which the probes hybridize) can interfere with hybridization, polymerase fill-in, and/or ligation. Furthermore, this may occur for only one allele, resulting in allelic drop-out, which ultimately decreases downstream sequencing accuracy. By using a set of MIPs having multiple hybridization sites for the capture of any given target, the probability of loss from polymorphism is substantially decreased because not all targeting arms in the set of MIPs will cover the location of the mutation.

Probes for MIP capture reactions are generally synthesized on programmable microarrays because of the large number of sequences required. Because of the low synthesis yields of these methods, a subsequent amplification step is required to produce sufficient probe for the MIP amplification reaction. The combination of multiplex oligonucleotide synthesis and pooled amplification results in uneven synthesis error rates and representational biases. By synthesizing multiple probe sequences for each target, variation from these sources should be averaged out because not all MIPs for a given target will have the same error rates and biases.

All five aspects of the invention involve using multiple MIPs (and therefore, multiple probes) to amplify each target. The set of probes for a given target can be designed to 'tile' across the target, capturing the target as a series of shorter sub-targets. Alternately, the set can be designed to 'stagger' the exact positions of the hybridization regions flanking the target, capturing the full target (and in some cases capturing flanking non-target sequence) with multiple MIPs having different targeting arms, obviating the need for tiling. The particular approach chosen will depend on the nature of the target set. For example, if small regions are to be captured, a staggered-end approach might be appropriate, whereas if longer regions are desired, tiling might be chosen. In all cases, the amount of bias-tolerance for pathological MIPs can be 'dialed in' by changing the number of different MIPs used to capture a given target.

The 'coverage factor', or number of MIPs used to capture a basepair in the target, is the most critical parameter to specify, and implies a different number of MIPs per target depending on whether one is using the tiling approach (Figure 1) or one of the staggered approaches (Figures 2, 3). In the tiling approach, for a desired coverage factor, the number of MIPs per target is a function of target length, sub-target length, and spacing between adjacent sub-target start locations. For example, for a desired coverage factor of 3, a 200 bp target with a start-site separation of 20 bp and sub-target length of 60 bp necessitates 10 MIPs (Figure 1). It follows that an arbitrary coverage factor can be achieved by varying the number of probes per target and the length of the subtargets captured. In the staggered approach, a fixed-length target is captured as several targets and super-targets of varying lengths. For example, a target of 50 bp can be captured at a coverage factor of 3 with 3 probes in either a 'staggered' (Figure 2) or 'alternating staggered' configuration (Figure 3).

ML-IPR-000012

Multiplex amplification strategies like MIPs are used either analytically, as in detection of SNPs, or preparatively, often for next-generation sequencing. In the preparative setting, the output of the MIP amplification reaction is generally the input to a shotgun library protocol, which then becomes the input to the sequencing platform. The shotgun library is necessary in part because next-generation sequencing yields reads significantly shorter than amplicons such as exons. In addition to the bias-reduction afforded by the multi-tiled approach described here, tiling also obviates the need for shotgun library preparation. Since the length of the sub-target can be specified when the MIPs are designed, it can be chosen to match the readlength of the sequencer. In this way, reads can 'walk' across an exon by virtue of the start position of each sub-target in the MIP set for that exon.

ML-IPR-000013

Claims

1. *A method for reducing bias (increasing uniformity) in a set of targets produced in a multiplex MIP reaction by designing multiple MIPs for each target*
2. *A method for increasing reproducibility of target abundance in a MIP reaction (both multiplex and uniplex) by designing multiple MIPs for each target*
3. *A method for reducing the effect of polymorphisms in flanking sequence on target abundance in a MIP reaction (both multiplex and uniplex) by designing multiple MIPs for each target*
4. *A method for reducing the effect of probe synthesis / probe amplification variability on uniformity/reproducibility in a MIP reaction (both multiplex and uniplex) by designing multiple MIPs for each target*
5. *A method for sequencing a target/set of targets captured in a MIP reaction without preparing a shotgun library by designing a set of MIPs for a target that tile across the target*

ML-IPR-000014

Figures

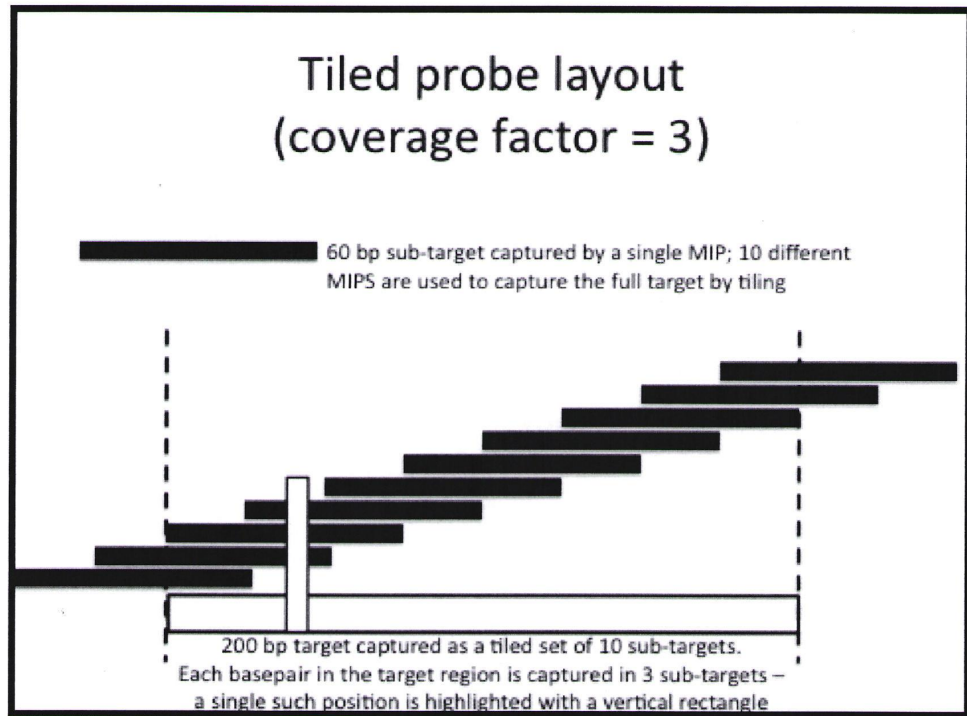


Figure 1. Tiled probe layout, showing ten captured sub-targets tiled across a single target. Each position in the target is covered by three sub-targets such that MIP performance per basepair is averaged across three probes.

ML-IPR-000015

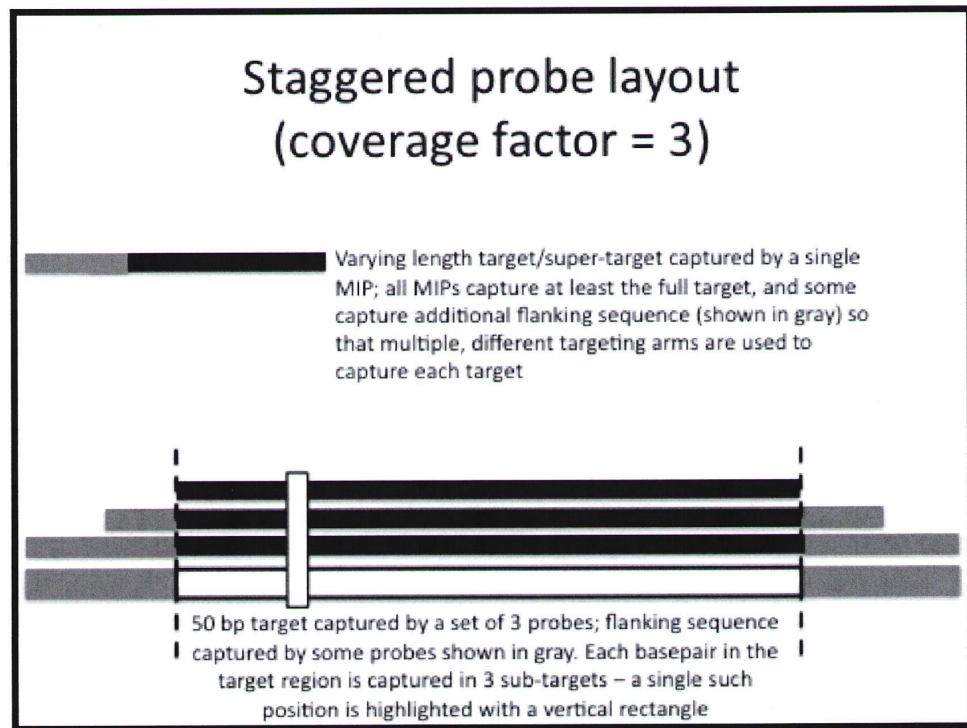


Figure 2. Staggered probe layout, showing the targets captured by a set of three MIPs. Each MIP captures the full target, shown in black, plus (in some cases) additional extra-target sequence, shown in gray, such that the targeting arms of each MIP fall on different sequence. Each position in the target is covered by three sub-targets such that MIP performance per basepair is averaged across three probes.

ML-IPR-000016

Alternating staggered probe layout (coverage factor = 3)

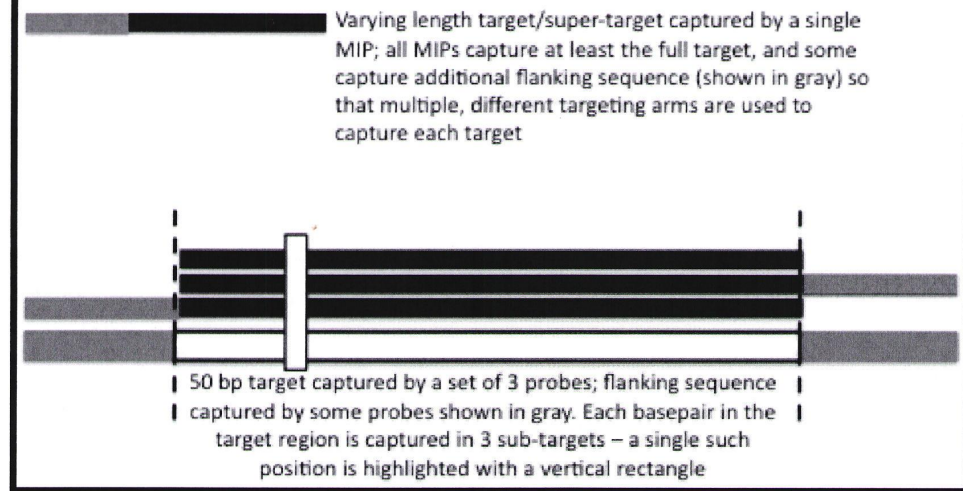


Figure 3. Alternating staggered probe layout, showing the targets captured by a set of three MIPs. Each MIP captures the full target, shown in black, plus (in some cases) additional extra-target sequence, shown in gray, such that the targeting arms of each MIP fall on different sequence. Each position in the target is covered by three sub-targets such that MIP performance per basepair is averaged across three probes.

ML-IPR-000017

Estimate of Target Length in Amplification-Based Assays

Drafted by Jin Billy Li, Ph.D., scientific advisor at Good Start Genetics, Inc.

Background:

1. Fragile X

From [*****en.wikipedia.org/wiki/Fragile_X_syndrome](https://en.wikipedia.org/wiki/Fragile_X_syndrome):

"Fragile X syndrome, or Martin-Bell syndrome, is a genetic syndrome which results in a spectrum of characteristic physical, intellectual, emotional and behavioural features which range from severe to mild in manifestation. The syndrome is associated with the expansion of a single trinucleotide gene sequence (CGG) on the X chromosome, and results in a failure to express the FMR-1 protein which is required for normal neural development. There are four generally accepted forms of Fragile X syndrome which relate to the length of the repeated CGG sequence; Normal (29-31 CGG repeats), Premutation (55-200 CGG repeats), Full Mutation (more than 200 CGG repeats), and Intermediate or Gray Zone Alleles (40 - 60 repeats)."

2. Generally, the capture efficiency is dependent on the target size.

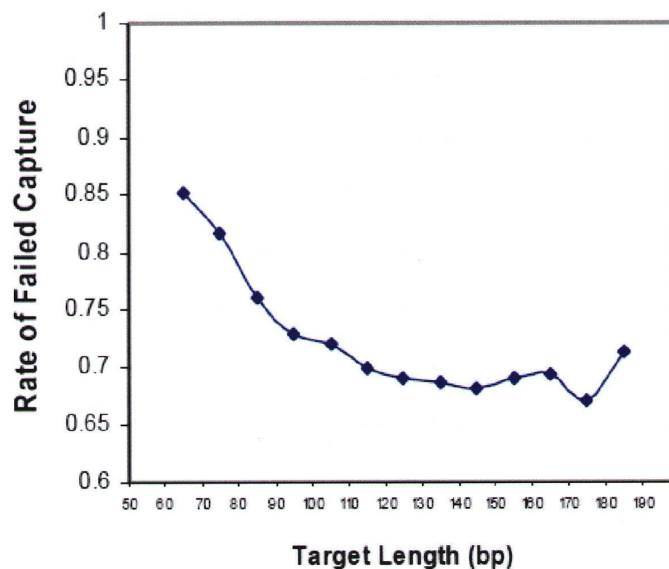


Figure 1. [Supplemental Fig. 3B from *Porreca et al (Nature Methods 2007)*: "The rate of failed capture is shown as a function of the length of the gap-fill sequence itself (range = 60 – 191 bp; ~10 bp bins). We observe that shorter gapfill lengths are associated with a modestly higher rate of 'dropout'."]

Description of invention:

This invention aims to estimate the size of genomic (or transcriptomic) regions that are hard to be accessed, such as repetitive elements. In this attempt, fragile X will be used as an example, where the size of trinucleotide repeats (genotype) is linked to the symptom (phenotype).

Repetitive regions are always surrounded by non-repetitive unique sequences, which can be used to amplify the repeat-containing regions using, for example, PCR or padlock (MIP)-based method.

In addition to the repetitive regions, the padlock probe is designed to include at least a sequence that is sufficient to be uniquely identified in the genome (or target pool). After the probe is circularized and amplified, the amplicon can be end-sequenced so that the unique sequence can be identified and served as the “representative” of the repetitive region (Fig. 3). Although capturing efficiency is overall positively correlated with target length, different probe sequences may have unique features. Therefore, multiple probes should be designed and tested so that the optimal one is chosen to be sensitive enough to differentiate repetitive sizes of roughly 0-150bp, 150-600bp, and beyond, which represent normal, premutation and full mutation of fragile X syndrome, respectively.

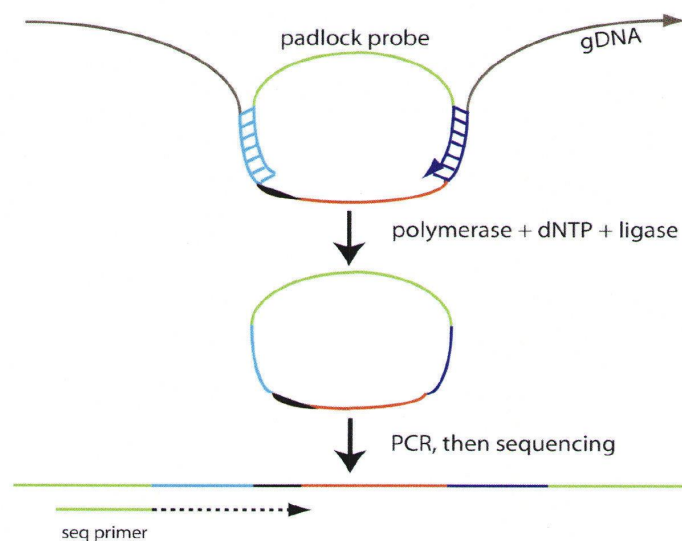


Figure 3. Scheme of padlock capture of a region that includes both repetitive regions (in red) and the adjacent unique sequence (in black). After the padlock is circularized and amplified, it can be end-sequenced to obtain the sequence of the unique sequence, which represents the repetitive region of interest.

The number of reads obtained for the “representative” of the repetitive region is not informative to estimate the target length because it is dependent on the total number of reads obtained. To overcome this, it is useful to include one or more probes that target other “control” regions where no or minimal polymorphism exists among populations. Because of the systematic consistency of capturing efficiency (Fig. 2), the ratio of reads obtained for the repetitive “representative” to reads obtained for the control region(s) will be tuned using DNA with defined numbers of repeats. Ultimately, the ratio can serve as a measure of the repeat length (Fig. 4).

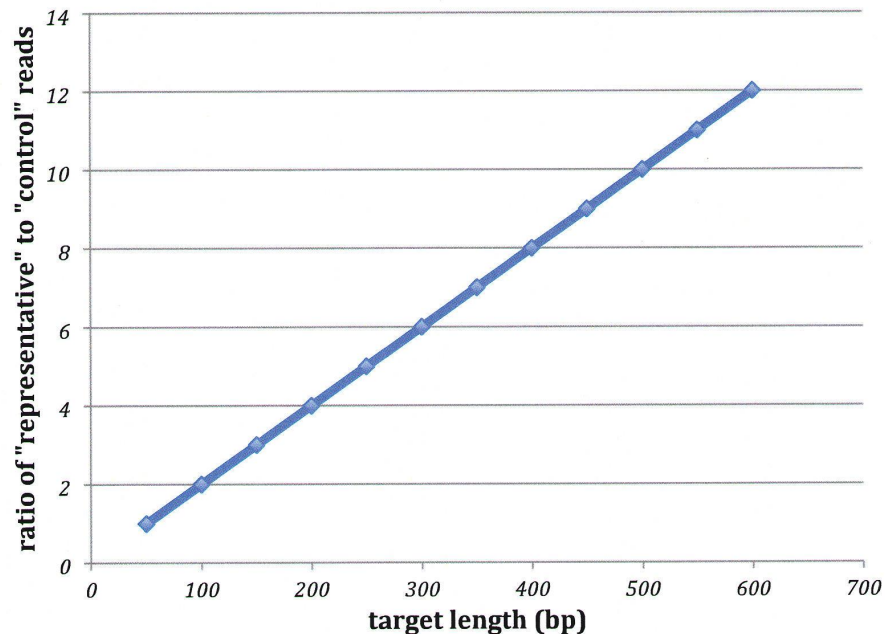


Figure 4. The hypothetical relationship between target gap size and the relative number of reads of the repetitive region, which is measured by the ratio of the repeat “representative” reads vs. the “control” region reads. The unit of y-axis is arbitrary.

To better tell targets with similar size range apart, the whole repetitive region can be sequenced by making a shotgun library. The longer the repeat is, the more short reads of repeats will be obtained. Therefore, the target length will contribute twice to the

relative number of “repetitive” reads, which will gainer better resolution of differentiating targets.