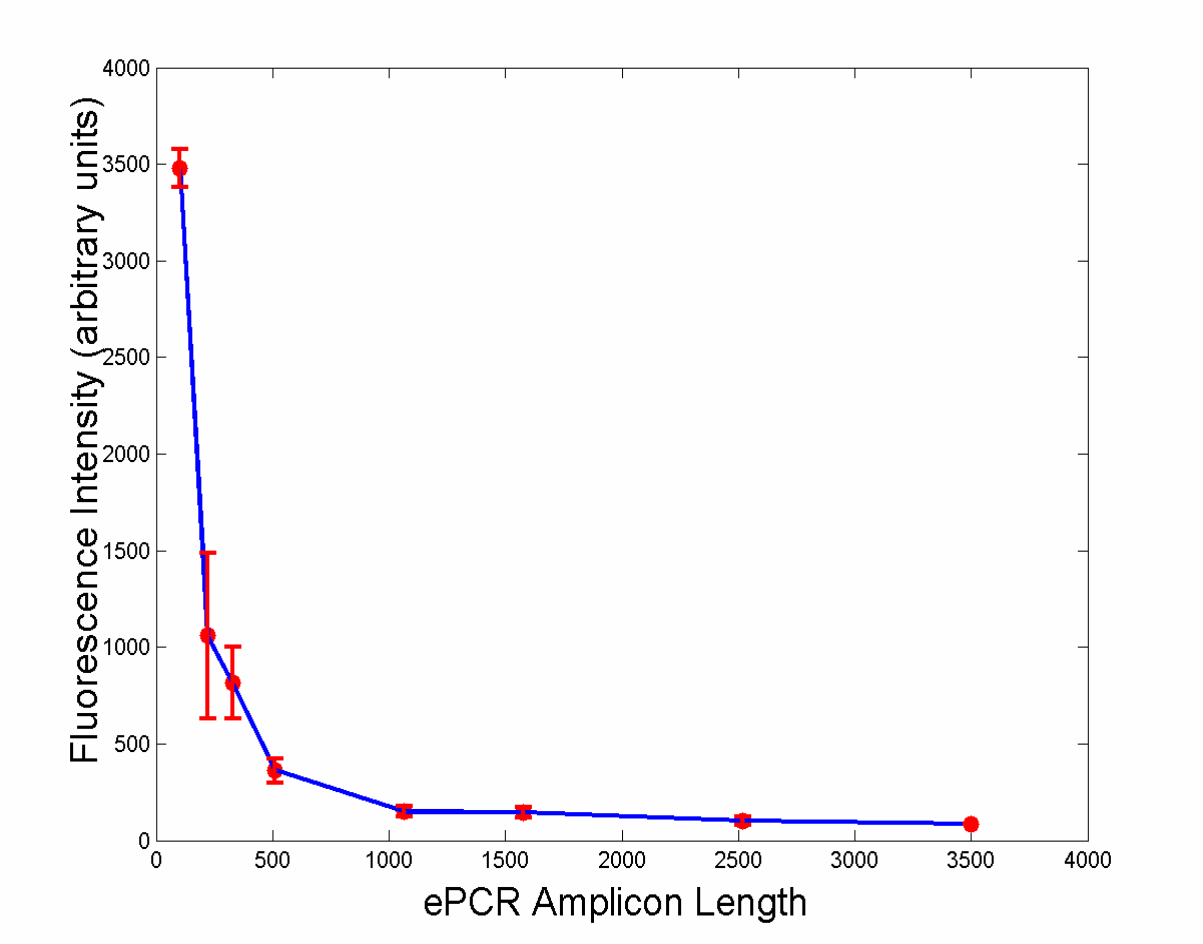
SUPPORTING ONLINE MATERIAL

Supplementary Fig. 1



Efficiency of Emulsion PCR Decreases Markedly With Increasing Amplicon Length. We investigated the efficiency of emulsion PCR (eCPR) as a function of amplicon length. The X-axis of the graph is amplicon length in base pairs, and the Y-axis of the graph represents the intensity of amplified beads in arbitrary units. Products of length 102, 220, 328, 510, 1064, 1575, 2519, and 3497 base-pairs were amplified in separate ePCR reactions. Beads from each reaction were monolayered in an acrylamide gel and an Alexa-546 labeled primer complementary to sequence present in all amplicons was hybridized to the beads. To quantify the efficiency of each emulsion PCR reaction, several frames of epifluorescence imaging (Cy3 channel) were performed. Beads meeting a minimum cutoff were selected (to identify amplified beads), and their average intensity was determined. For each template length, three experiment replicates were carried out.

Experimental details are as follows. Note that the deviations from protocols of Dressmann et al. (2003) described here are specific to this experiment and differ from optimizations described elsewhere in the manuscript. The same caveat applies to the details of gel-pouring, hybridization, and imaging. We suggest using the detailed protocols provided in Supplementary Notes 2-8 rather than the those used for this experiment. In the nomenclature used here and

elsewhere in the manuscript, the primer that is biotin-immobilized to the magnetic beads is referred to as the forward primer (and the other as the reverse primer).

Emulsion PCR was performed as described Dressmann et al. (2003) with the following differences: (a) The PCR conditions were 1x Platinum Taq PCR buffer (Invitrogen), 6.5 mM $MgCl_2$, 375 µM dNTPs, 50 nM 5 µM free forward primer (JMP130), $5x10^8$ template molecules, 40 Units Platinum Taq (Invitrogen). The PCR thermocycling conditions were 94°C for 3 min, 10 cycles of 94°C for 45 sec, 65°C for 30 sec, 72°C for 3.5 min, and 50 cycles of 94°C for 45 sec, 57°C for 30 sec, 72°C for 3.5 min. After emulsions were broken, beads were washed 2x in TE and incubated in 30ul of 0.1 M NaOH for 5 minutes at room temperature to remove the free strand from amplified beads. Stripping solution was neutralized by the addition of 50ul of 1 M Tris (pH 7.5). Beads were washed 2x in 50ul TE and resuspended in 5ul of TE. 2ul of beads were poured into 20% acrylamide, 0.25% bisacrylamide gels on standard polony slides (ER-285W, Erie Scientific, Portsmouth, NH). Gels were polymerized at 4°C to allow for monolayering of beads. Polymerized gels were washed once in water for 2 minutes and once in Wash 1E (10 mM Tris [pH 7.5], 50 mM KCl, 2 mM EDTA [pH 8.0], 0.01% Triton X-100) for 5 minutes. Hybridization was done in 6x SSPE with 0.5 uM of the Cy3-labeled oligonucleotide (JMP109), heated to 94°C for 6 minutes, cooled to 55°C for 1 minute, and then washed 2x for 5 minutes in Wash 1E at room temperature.

Emulsion PCR beads were generated for amplicons of length 102, 220, 328, 510, 1064, 1575, 2519, and 3497 base-pairs in separate reactions, in triplicate. The same template and the same beads, bearing the same forward primer, were used for all amplicon lengths. The amplicon lengths were thus determined by changing the reverse primer. A negative control experiment contained no reverse primer.

The beads were visualized on a Nikon Eclipse TE2000-U scope with an ExFo X-Cite120 lamp and a 30x objective. Two 16-bit, 1392x1040 images were captured for each field – one a 3 ms exposure in brightfield and the second a 250 ms exposure with epifluorescence illumination and a green Cy3 exciter filter. With MATLAB scripts, bead locations were determined from the brightfield image, and bead intensities were calculated by averaging a 3x3 window around the centroid of each bead.

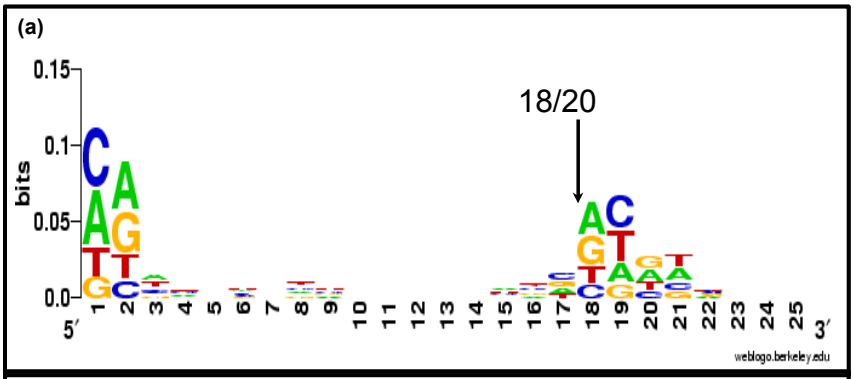
For each slide, two fields of view were chosen at random (five for amplicon lengths of 220 bp and 328 bp). For each field, the average intensity of amplified beads was calculated. Amplified beads were identified as those with an average intensity greater than 70 (except for the negative control experiment, in which all beads were considered included in the calculation). The values for each field were then averaged to determine an intensity for a given amplicon length. For each amplicon length, the mean and standard deviation of three replicates are plotted above.

Template: A 3801 bp linear PCR product amplified from *E.coli* 16S rDNA cloned into the pCRII-TOPO vector (Invitrogen).

Primer Sequences:

JMP109	Hybridization Primer	Alexa546-CTGAGCCAKGATCAAACTCT
JMP127	Bead Immobilized Forward Primer	(Dual Biotin)-TTTTTTTATGACCATGATTACGCCAAGCTATTTAGGTGACA
JMP130	Free Forward Primer	CCATGATTACGCCAAGCTATTTAGGTGACA
JMP184	Reverse Primer for 102 bp Amplicon	TTCTGAGCCATGATCAAACTCTTCGGTACCAAGCTTGATGCATAG
JMP185	Reverse Primer for 220 bp Amplicon	AAGCTTCTTCCTGTTACCGTTCGAC
JMP186	Reverse Primer for 328 bp Amplicon	GTCTTGCGACGTTATGCGGTATTAG
JMP187	Reverse Primer for 510 bp Amplicon	ATTGTGCAATATTCCCCACTGCTG
JMP188	Reverse Primer for 1064 bp Amplicon	CCCCCGTCAATTCATTTGAGTTTTA
JMP189	Reverse Primer for 1575 bp Amplicon	ACCTACTTCTTTTGCAACCCACTCC
JMP190	Reverse Primer for 2519 bp Amplicon	CTGCAAGCTACCTGCTTTCTCTTTG
JMP191	Reverse Primer for 3497 bp Amplicon	AACTCGTCAAGAAGGCGATAGAAGG

Supplementary Fig. 2



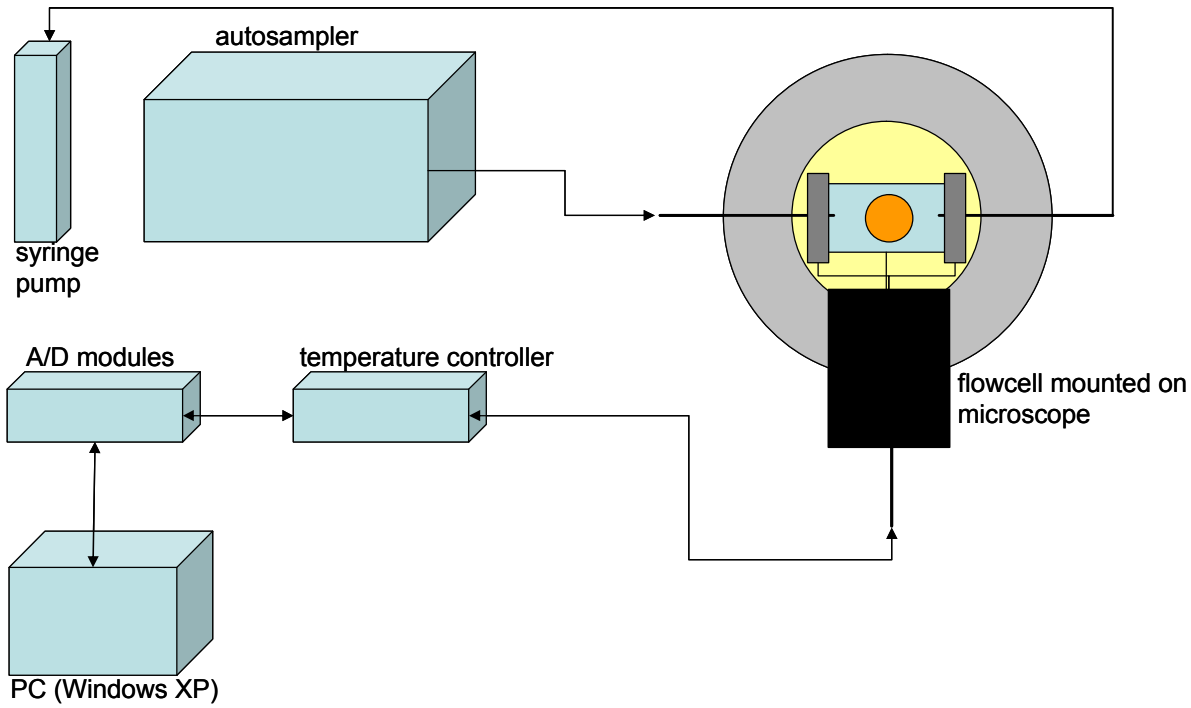
(c)

		Tag Position			
		1	2	18	19
17 bp tags	G	0.15	0.29	0.15	0.32
	A	0.37	0.36	0.20	0.32
	C	0.31	0.15	0.35	0.17
	T	0.17	0.20	0.31	0.19
18 bp tags	G	0.13	0.31	0.31	0.16
	A	0.34	0.36	0.34	0.19
	C	0.35	0.15	0.17	0.33
	T	0.18	0.19	0.19	0.32
combined	G	0.14	0.30	0.23	0.23
	A	0.36	0.36	0.28	0.26
	C	0.33	0.15	0.25	0.25
	T	0.17	0.19	0.24	0.26

Junctional Biases in the Paired Tag Library. The *in vitro* library construction procedure contains several enzymatic steps (e.g. A-tailing, circularization, MmeI digestion), raising the question of whether sequence context will bias tag generation. We therefore asked whether deviations from background nucleotide frequency were observed across the length of the tags. Each tag has two junctions, one derived from circularization of the A-tailed genomic fragment with the T-tailed T30 linker, and the other derived from MmeI digestion, end-repair, and ligation of flanking primer oligonucleotides (Supp Note 1). At roughly a 1:1 ratio, MmeI digests at a distance of 18/20 or 19/21 from its recognition site (Dunn, et al. 2002). In our protocol these yield 17 bp or 18 bp tags of unique genomic sequence, respectively. As sequence context may influence the length of the tag that is cut by MmeI, we examined 17 bp and 18 bp tags separately. By looking at tags that have been mapped back to the genome, we are also able to measure the influence of sequence context beyond the tag itself. Logos were generated by WebLogo (Crooks 2004) using 10,000 mapped 17 bp tags (panel a) or 10,000 mapped 18 bp tags (panel b). The height of the logo stacks is proportional to the information content at each position, reflecting

deviations from background nucleotide frequency. In the numbering scheme of the X-axes, the five bp MmeI recognition site (TCCRAC) would be positioned at -6..-1, followed by an A (from the A-tailing of sheared genomic fragments) at position 0. Deviations from background nucleotide frequency are seen at both the circularization junction and at the primer junction. The nucleotide frequencies at the four most biased positions (1, 2, 18 and 19) are listed in panel c. The biases observed at positions 1 and 2 are probably due to context-dependent efficiency of either the A-tailing or TA ligation steps. The modest biases at positions 18 and 19 are reciprocal in that they are only seen when the tags are segregated by length, indicating that the immediate sequence context of the cut site influences, but does not determine, the choice by MmeI to cut 18/20 or 19/21.

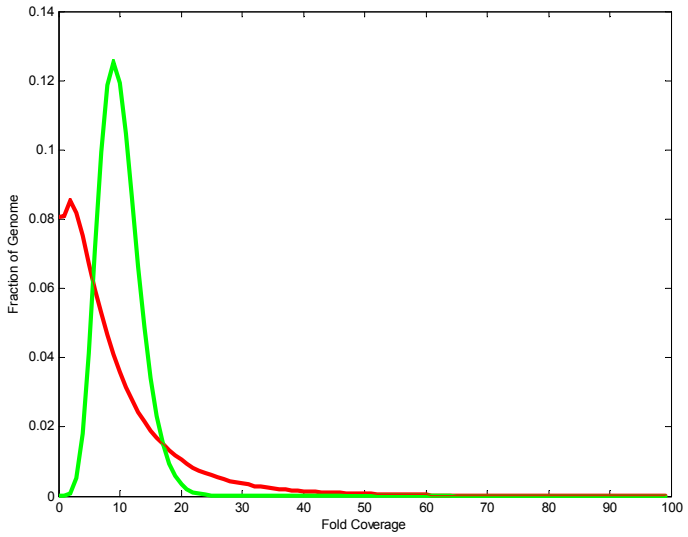
Supplementary Fig. 3



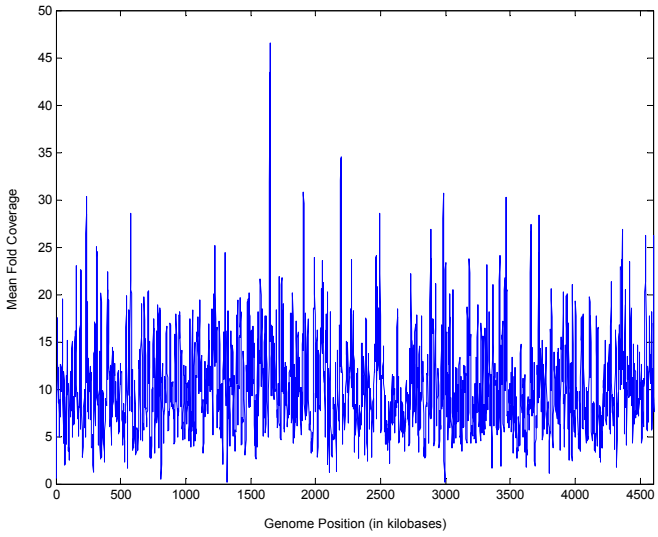
Schematic of a Automated Sequencing-By-Imaging Platform. Autosampler (Alcott Chromatography) allows any of up to 96 different solutions to be applied to the sequencing array mounted in the flowcell (shown in orange). Incubation temperature is under computer control (Superlogics) and can be varied from 25°C to 60°C. Custom flowcell (Bioptechs) is attached directly to the microscope stage (Prior) and remains in fixed position in X, Y, and Z during the course of a sequencing run.

Supplementary Fig. 4

(a)



(b)



Distribution of Base Calls Across *E. coli* Genome. (a) The X-axis represents specific levels of coverage, i.e. 0x, 1x, 2x, etc., and the Y-axis represents the fraction of the reference genome at this fold coverage. In red, a histogram based on the collected sequencing data. In green, the expectation based on the Poisson distribution, i.e. assuming a completely unbiased library (green). (b) Histogram of mean fold-coverage across the 4.64 Mb genome with ~48.2 Mb of sequence. Windows are 5 kb, with points calculated every 1 kb.

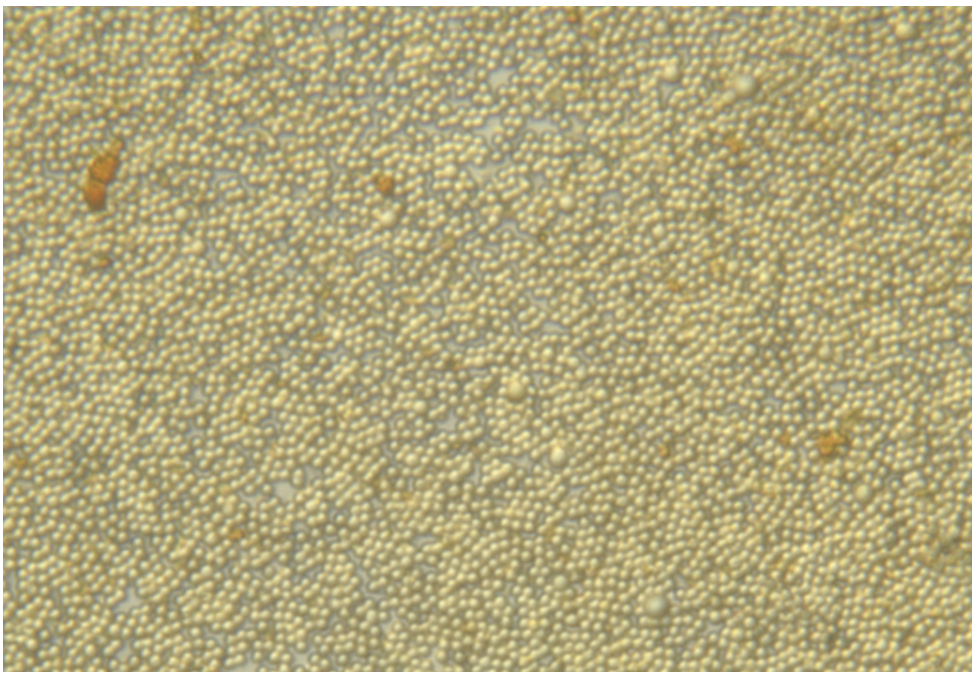
Supplementary Fig. 5

	Zero	Unique	Multiple
Paired, no substitutions	0.0%	**90.4%**	9.7%
Paired, one substitution	0.0%	**92.8%**	7.2%
Unpaired, no substitutions	96.0%	1.5%	2.5%

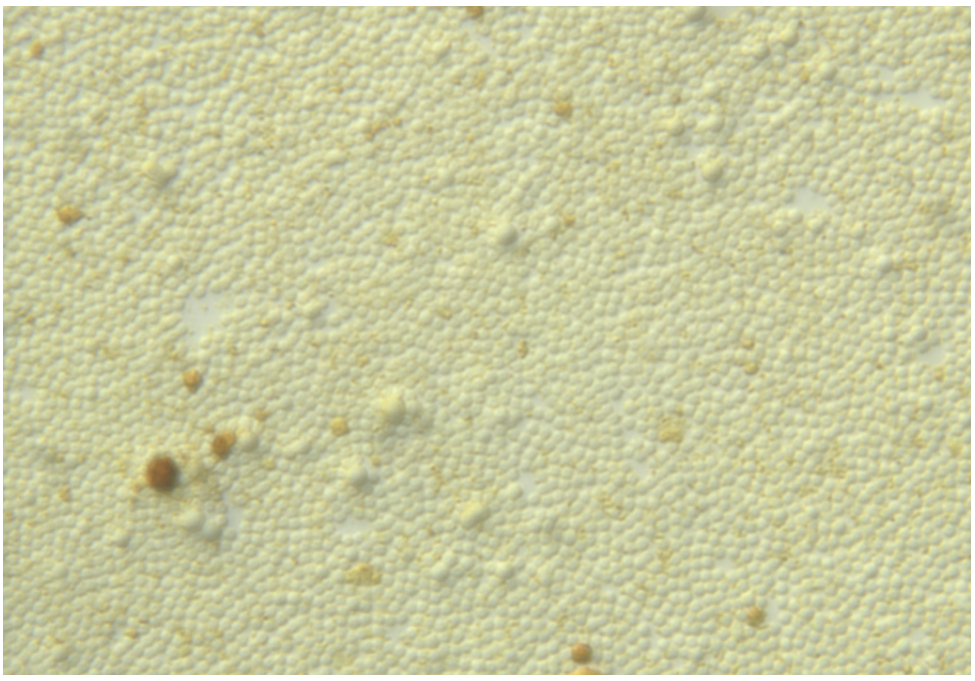
Simulation of Human Genome Resequencing. Simulations were performed to determine whether mate-paired 17 bp tags could reasonably be applied to human genome resequencing. Sets of 10,000 mate-paired tag-sets (17 contiguous bases per each of two tags) separated by 750 – 1150 bp were derived *in silico* from a single human 40 megabase contig. A control set of 10,000 unpaired tag-sets was also generated to determine the rate of spurious matching. Each set was matched against the human genome using algorithms identical to those used for the *E. coli* resequencing data, i.e. allowing for a single substitution per read-pair. We asked whether tags could be matched to zero, one, or more than one location with results shown above. Over 90% of mate-paired 17 bp tag-sets could be matched to a unique location

Supplementary Fig. 6

A



B



Self-organized monolayers of 1.3 micron streptavidin coated beads formed using (A) SDS or (B) Tween as a detergent.

Materials:

Teflon Rings. Our teflon rings have a circular cross section and can be found at most hardware stores (ours were purchased at Lowe's). These rings have a 5.5mm inner diameter and a 11mm outer diameter.

Beads. 1.31uM streptavidin coated beads (Spherotech).

Protocol:

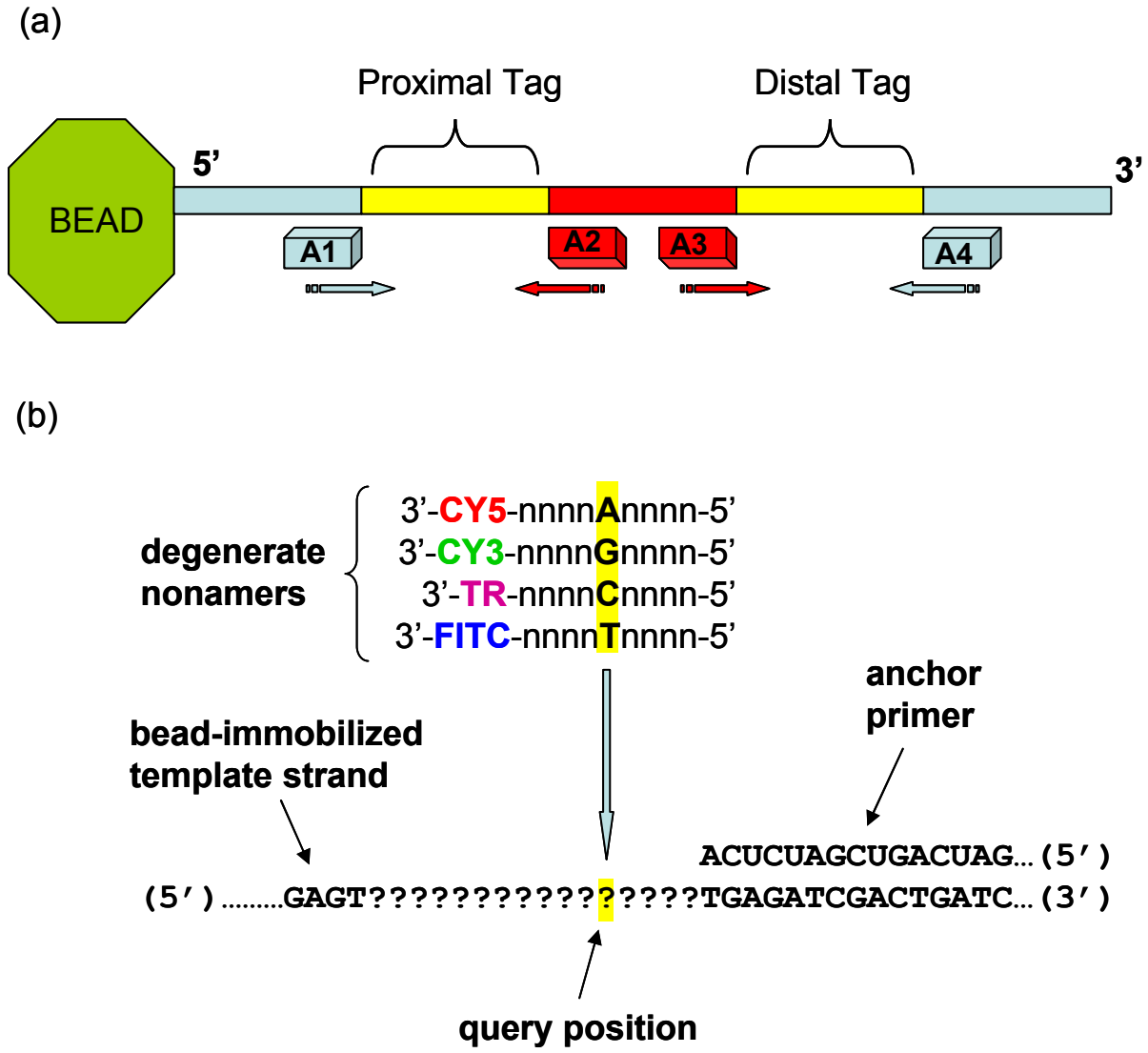
4.2e7 beads are suspended in 19.2 microliters TE + detergent (10mM SDS or 100uM Tween). The Teflon rings are clamped to bind-silane treated, teflon coated slides (Erie Scientific). The bead slurry is pipetted into the center of the oval ring and the liquid is allowed to evaporate in a fume hood with laminar flow. After drying, the ring is removed and a coverslip is placed over the oval. A 6% acrylamide gel is then poured over the beads by pipetting the mixture under the coverslip, covering the beads.

This protocol for producing monolayers is a modified from the following papers:

Denkov N.D., Velev O.D., Kralchevsky P.A . Ivanov I.B., Yoshimura H, and Nagayama. K. Mechanism of Formation of 2-Dimensional Crystals From Latex-Particles on Substrates, *Langmuir*, vol. 8, no. 12, pp. 3183-3190.

Gigault C., Dalnoki-Veress K.,and Dutcher J.R. Changes in the Morphology of Self-Assembled Polystyrene Microsphere Monolayers Produced by Annealing, *Journal of Colloid and Interface Science*, vol. 243, no. 1, pp. 143-155.

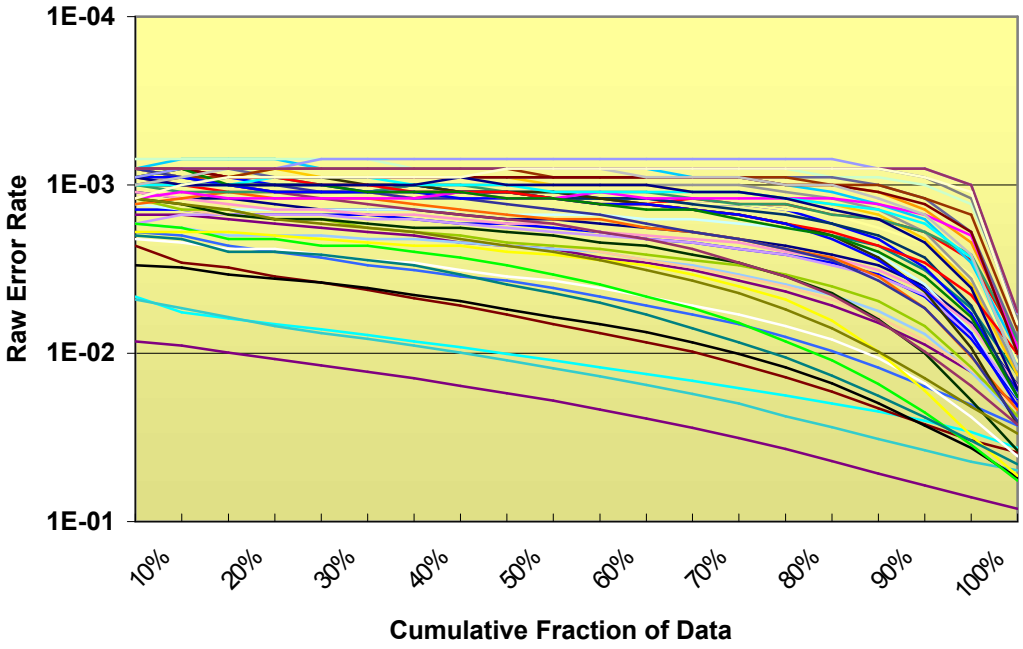
Supplmentary Fig. 7



Sequencing By Ligation. The structure of sequencing cycles is described in the text and Supp Note 6. In brief, cycles consist of the following four steps: (a) hybridization of anchor primer, (b) ligation of fluorescent, degenerate nonamers, (c) four color imaging on epifluorescence microscope, (d) stripping of the anchor primer:nonamer complexes prior to beginning the next cycle. The images above are intended to clarify the positioning of anchor primers and the nonamer ligation steps. (a) Hybridization of one of four "anchor primers". The anchor primers are each designed to be complementary to universal sequence immediately 5' or 3' to one of the two tags. A1, A2, A3 and A4, as shown above, indicate the four locations to which anchor primers are targeted relative to the amplicon. Arrows indicate the direction sequenced into the tag from each anchor primer. From anchor primers A1 and A3, we sequence 7 bases into each tag, and from anchor primers A2 and A4, we sequence 6 bases into each tag. Thus, we obtain 13 bp per tag, and 26 bp per amplicon, with 4 to 5 bp gaps within each tag sequence.

(b) **Ligation of fluorescent, degenerate nonamers.** Each cycle involves performing a ligation reaction with T4 DNA ligase and a fully degenerate population of nonamers. The nonamer molecules are individually labeled with one of four fluorophores (e.g. Texas Red, Cy5, Cy3, FITC). Depending on which position that a given cycle is aiming to interrogate, the nonamers are structured differently. Specifically, a single position within each nonamer is correlated with the identity of the fluorophore with which it is labeled. Additionally, the fluorphore molecule is attached at the opposite end of the nonamer relative to the end targeted to the ligation junction. For example, in the graphic shown here, the anchor primer is hybridized such that its 3' end is adjacent to the genomic tag. To query a position five bases in to the tag sequence, we use the four-color population of nonamers shown here.

Supplementary Fig. 8



Cumulative Raw Accuracy By Cycle. Cumulative distribution of raw error as a function of rank-ordered quality, with each of the sequencing-by-ligation cycles in both the 18.1 Mb and the 30.1 Mb experiments considered as an independent dataset. The *x*-axis indicates percentile bins of beads, sorted on the basis of a confidence metric. The *y*-axis (log scale) indicates the raw base-calling accuracy of each cumulative bin.

SUPPLEMENTAL NOTE 1
Library Protocol

The library construction protocol takes approximately 1 week and has the following steps:
(a) purification of genomic DNA
(b) shearing of genomic DNA to generate fragments
(c) end-repair and A-tailing of DNA fragments
(d) PAGE size-selection of sheared fragments
(e) circularization with T-tailed spacer oligonucleotide ("T30")
(f) rolling circle amplification (RCA) with random hexamers
(g) digestion with MmeI (type IIs) to release paired tags
(h) PAGE purification of tag-T30-tag library
(i) end-repair of tag-T30-tag library
(j) ligation of FDV2 (PR1-F) and RDV2 (PR1-R) primer oligonucleotides
(k) PAGE size-selection of paired-tag library
(l) nick translation to eliminate nicks in dsDNA library
(m) PCR amplification of paired-tag library
(n) PAGE size-selection of paired-tag library
(o) Library validation via cloning and Sanger sequencing

Schematic of each library molecule:

(a) Purification of genomic DNA
For each strain, cultures were grown overnight in 3 ml of LB and isolated with the Qiagen DNeasy Tissue kit as per the manufacturer's protocol. Yield for genomic DNA purification was approximately 30 ug.
http://www1.qiagen.com/literature/handbooks/PDF/GenomicDNAStabilizationAndPurification/FromAnimalAndPlantIssues/DNY_Tissue_Kit/1026641HBDNY_0304WW_LR.pdf

(b) Shearing of genomic DNA to generate fragments with a broad size distribution
In the case of these libraries, shearing of genomic DNA was kindly performed for us by Agencourt Biosciences.
http://www.genomicsolutions.com/files/HydroShear.pdf
The size-distribution of the resultant DNA fragments was quite broad, as can be seen on the gel below

(c) End-repair and A-tailing of DNA fragments
Unless stated otherwise, all DNA quantitation was performed on a Nanodrop ND-1000 Spectrophotometer.
Quantified sheared genomic DNA: 57 ng/uL.

Sheared DNA fragments were end-repaired with the EpiCentre End-It DNA End Repair Kit:
http://www.epicentre.com/pdftechlit/153pl053.pdf

- 170ul of sheared E.coli DNA (~9-10 ug)
- 25ul of 10x EndIt Buffer
- 25ul of 10x EndIt ATP
- 25ul of 10x EndIt dNTPs
- 5ul of EndIt Enzyme
- Total volume = 250 ul.
- Reaction was incubated at room temperature for 1 hour.

DNA was purified on a Qiagen Qiaquick column as per manufacturer's recommendations for PCR product purification.
http://www1.qiagen.com/literature/handbooks/PDF/DNACleanupAndConcentration/QQ_Spin/1021422_HBQQSpin_072002WW.pdf

- ~90 ul of Buffer EB was used for elution

- DNA quantitated at 96.5 ng/uL

Volume was split to 4 tubes of ~22 ul.

To eliminate residual enzyme activity, tubes were heated to 70'C for 15 minutes.

An A-tailing master-mix was prepared as follows:

- 100 ul of 10x PCR buffer (no MgCl2) [Invitrogen]
- 60 ul of 50 mM MgCl2 [Invitrogen] -> final concentration of 3 mM
- 5 ul of 100 mM dATP [Invitrogen] -> final concentration = 0.5 mM
- 5 ul of Taq (5 U/uL) [NEB]
- 610 ul of dH20

After heat-inactivation, added 78 ul of this mix to each tube containing 22 ul of sheared, end-repaired DNA fragments.

Tubes were incubated at 70'C for 30 minutes in a thermal cycling machine. The cycling program ended by cooling the tubes to 4'C, and they were transferred from the thermal cycler directly to ice.

DNA was purified by phenol-chloroform extraction and ethanol precipitation ("P:C:P"). Both here and elsewhere in the protocol, this was performed as follows:

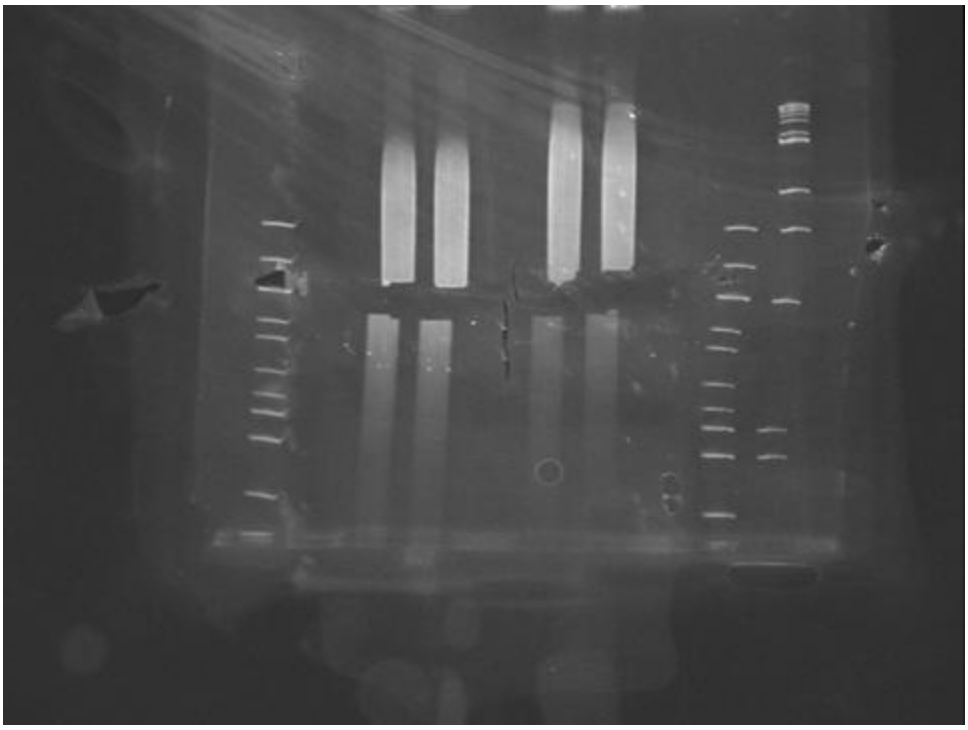
- Add an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1)
- Add 0.1 volumes of 3M NaOAc (pH = 5.2)
- Add 1.0 ul of glycogen (20 ug/uL)
- Add 2.5 volumes of cold 100% ethanol (from bottle stored at -20'C)
- Mix by inverting
- Put tube at -70'C for ~30-60 minutes
- Spin at maximum speed on microcentrifuge in 4'C room for 10 minutes
- Remove supernatant
- Add 1 ml of 80% ethanol
- Spin at maximum speed on microcentrifuge at room temperature for 5 minutes
- Remove supernatant
- Place tube on Speed-Vac for ~5 minutes
- Resuspend pellet in Buffer EB (Qiagen) or TE

In this case, each pellet was resuspended in 40 ul of Buffer EB.

(d) PAGE size-selection of sheared fragments

Loaded half of the material to a pre-cast 6% TBE PAGE gel (Invitrogen). 20 ul of DNA was mixed with 5 ul of 5x High-Density Sample Buffer (Novex). This same loading buffer is used for all subsequent PAGE gels in this protocol 12.5 ul of the sample/loading buffer mixture was loaded per lane (so two lanes per library). Gel was run on standard apparatus and a region corresponding to ~1000 bp fragments was cut out with minimal exposure to UV.

PAGE size-selection of sheared fragments, post-cutting.



1 100 bp ladder (0.5 ul of NEB 100 bp ladder)
2
3 “M” strain DNA
4 “M” strain DNA
5
6 Reference strain DNA
7 Reference strain DNA
8
9 100 bp ladder (0.5 ul of NEB 100 bp ladder)
10 1 kb ladder (0.5 ul of Invitrogen 1 kb ladder)

Gel fragments were diced with razor transferred to 600 ul of PAGE elution buffer [10 mM Tris-HCl (pH 7.5), 50 mM NaCl, 1 mM EDTA (pH 8.0)]. Tubes were put at 37’C overnight.
Next day, spun down elutions (1 minute at maximum speed on microcentrifuge) and transferred supernatants to new tubes. To improve recovery, washed gel fragments with an additional 200 ul of PAGE buffer. DNA was purified by P:C:P protocol as described in step (b). Resuspended each pellet in 22 ul of Buffer EB.

We were concerned that partial degredation of the A-tails may have occurred during the gel purification. We therefore chose to repeat the A-tailing step. In future iterations of this protocol, we feel it probably makes more sense to perform both end-repair and A-tailing after the PAGE-based size-selection, rather than before.

An A-tailing master-mix was prepared as follows:

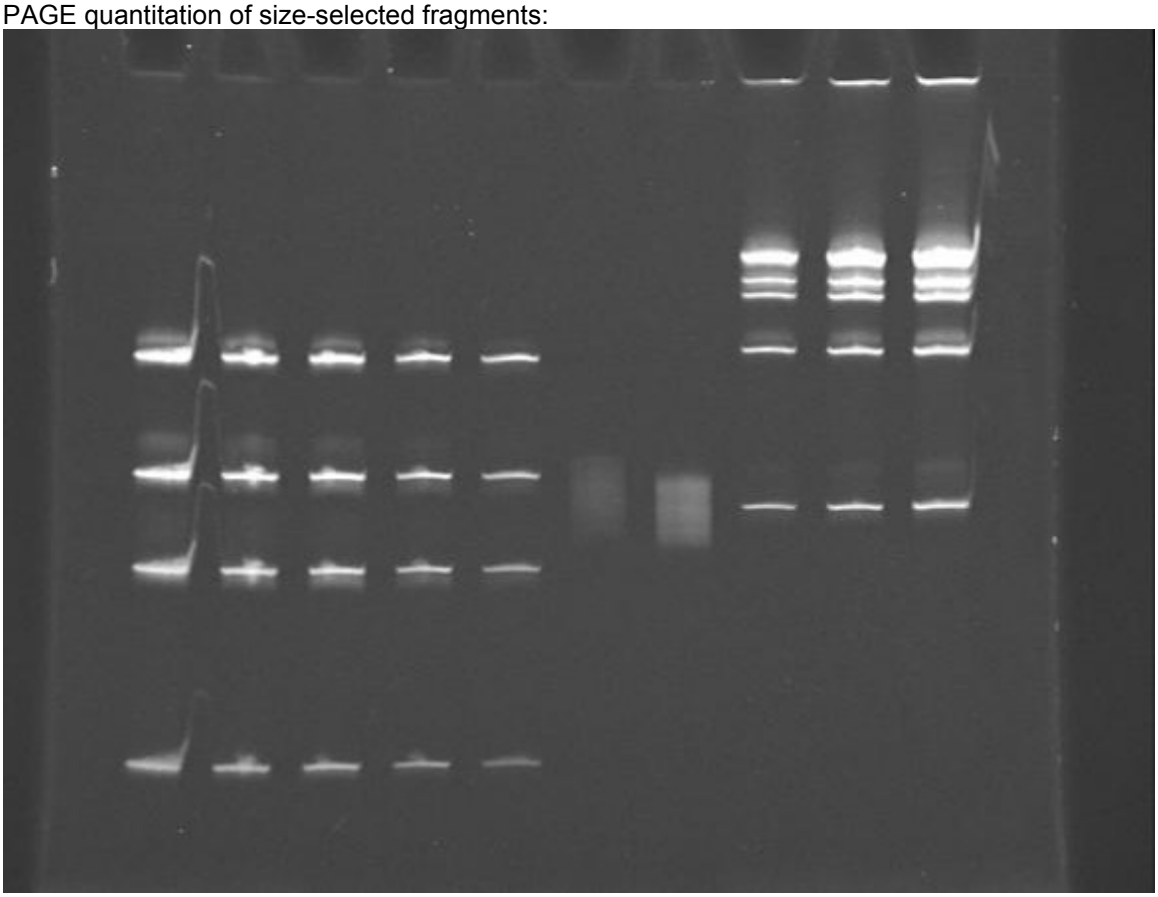
- 25.00 ul of 10x PCR buffer (no MgCl2) [Invitrogen]
- 15.00 ul of 50 mM MgCl2 [Invitrogen] -> final concentration of 3 mM
- 1.25 ul of 100 mM dATP [Invitrogen] -> final concentration = 0.5 mM
- 1.25 ul of Taq (5 U/uL) [NEB]
- 152.50 ul of dH20

Added 78 ul of this mix to each tube containing DNA resuspended in 22 ul for total volume of 100 ul. This was split to two thermal-cycler compatible tubes of 50 ul each.
Tubes were incubated at 70’C for 30 minutes in a thermal cycling machine.

The cycling program ended by cooling the tubes to 4'C, and they were transferred from the thermal cycler directly to ice.
DNA was purified by P:C:P protocol as described in step (b). Resuspended each library in 10 ul of Buffer EB and put on ice.

To quantitate recovery and estimate the size-range of recovered fragments, we ran a pre-cast 6% TBE PAGE gel (Invitrogen) using 20% of the purified material.

PAGE quantitation of size-selected fragments:



1 8.0 ul of High Mass DNA Ladder [NEB]
2 4.0 ul of High Mass DNA Ladder [NEB]
3 2.0 ul of High Mass DNA Ladder [NEB]
4 2.0 ul of library
5 2.0 ul of reference
6 0.5 ul of Low Mass DNA Ladder [NEB]
7 1.0 ul of Low Mass DNA Ladder [NEB]
8 2.0 ul of Low Mass DNA Ladder [NEB]
9 4.0 ul of Low Mass DNA Ladder [NEB]
10 8.0 ul of Low Mass DNA Ladder [NEB]

Gel-based quantitation of library = 43 ng (in 2 ul, or 20% of what we had); range = ~850-1150, mean = ~1000
Gel-based quantitation of reference = 18 ng (in 2 ul, or 20% of what we had); range = ~900-1250; mean = ~1075
So in the remaining ~8 ul volume that we have for each library, we have ~171 ng of library fragments and ~73 ng of reference fragments.

(e) Circularization with T-tailed spacer oligonucleotide ("T30")

Here we are planning to circularize the A-tailed library fragments with the T-tailed spacer oligonucleotide, which we are calling "T30". The T30 segment was prepared by annealing two 32-bp oligonucleotides to generate a 30 bp dsDNA fragment with single base "T" overhangs:

5' GTCGGAGGCCAAGGCGGCCGTACGTCCAACT 3'
3' TCAGCCTCCGGTTCCGCCGGCATGCAGGTTG 5'

Note that the T30 segment is flanked by outward-facing MmeI sites, and each 5' end is phosphorylated.

Annealing of the oligos was performed by mixing to a final concentration of 50 uM for each oligo, heating to 95'C for 10 minutes in a thermal cycler, shutting the thermal cycler off and allowing the mixture to cool slowly back to room temperature over the course of an hour.

The actual ligation of the T30 fragment with the A-tailed library fragments was performed with NEB's Quick Ligation kit:
http://www.neb.com/nebecomm/products/productM2200.asp
Reactions were prepared as follows.

- 8.0 ul of A-tailed library fragments (~171 ng @ 1000 bp = 0.2599 picomols)
- 27.2 ul of dH20
- 0.8 ul of T30 (1 uM starting concentration -> 0.8 pmols; 3-fold molar excess)
- 40.0 ul of 2x Quick Ligation Buffer
- 4.0 ul of Quick T4 DNA Ligase

Mix the reaction well before and after adding enzyme to each tube.
Incubated reaction for 10 minutes at room temperature, then moved to ice.
Heat-inactivated ligase on thermal cycler (65'C for 10 minutes)

To destroy all uncircularized material, we add an exonuclease mix. The exonuclease mix is prepared as follows:

- 4.0 ul of Exonuclease I (NEB, 20U/uL)
- 0.4 ul of Exonuclease III (NEB, 100U/uL)
- 35.6 ul of TE

To the 80 ul library reaction, add 10 ul of exonuclease mix. Incubate tube for 45 minutes on thermal cycler at 37'C, followed by 80'C for 20 minutes to heat-inactivate the exonucleases. This material is used directly in the RCA reaction of the next step, rather than going through any purification.

(f) Rolling circle amplification (RCA) with random hexamers
Here we are performing hyperbranched RCA to amplify the amount of library material that we have to work with. We used the RepliPhi phi29 kit from Epicentre.
http://www.epicentre.com/pdftechlit/201pl044.pdf
A master-mix was prepared as follows:

- 32.0 ul of dNTP mix (25 uM each)
- 80.0 ul of 10x RepliPhi phi29 reaction buffer
- 40.0 ul of random DNA hexamers (1 mM, synthesized as 5'-NNNN*N*N-3', where "*" indicates phosphorthioate linkage)
- 552.0 ul of dH20
- 16.0 ul of 5x SybrGreen

30ul library material was added to 270ul master-mix, for a total volume of 300 ul.. Split to 6 tubes of 50 ul. To denature circularized template, heat tubes to 95'C for 5 minutes, followed by rapid cooling to 4'C. Add 2.5 ul of phi29 enzyme to each tube on ice (for total volume of 52.5 ul per tube). Mix well, keeping on ice. Incubate tubes overnight at 30'C on thermal cycler.
We ran these RCA reactions on a real-time PCR machine, and were therefore able to monitor amplification via the SybrGreen dye included in the reaction. The dsDNA content had risen and leveled off by the 2-hour time-point, suggesting that running the reaction overnight may not be necessary.

DNA was purified with a Millipore Microcon-30 column, washing with a total of 1 ml of TE. The pellet was substantial- several washings of the Microcon-30 membrane are suggested to maximize recovery. A

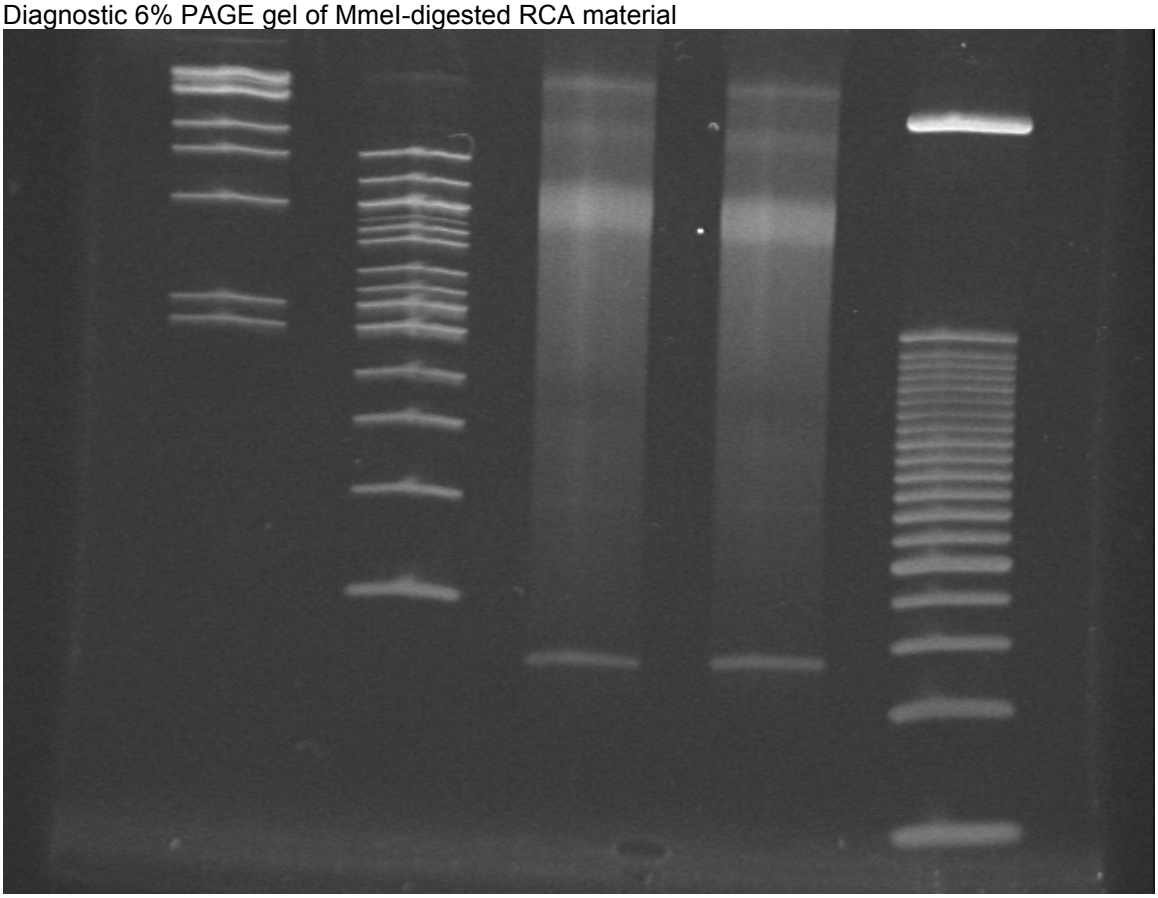
combination of heating (~50'C) and adding additional resuspension buffer (Buffer EB) was used to resuspend the DNA. Ended up with approximately 750 ul of library.
Quantitated samples on Nanodrop: 230 ng/uL. The RCA reaction thus resulted in ~150 ug of library.

(g) Digestion with MmeI (type IIs) to release paired tags
Digested ~40 ug of library with MmeI. As the MmeI site cuts at a distance from its recognition site in the T30 segment, and there are outward-facing MmeI sites at either end of the T30 segment, this digestion is expected to release the T30 segment flanked by ~18 bp tags with 2 bp overhangs (~70 bp in length). Because genomic fragments were circularized with T30 prior to MmeI digestion, these tags are expected to be paired with respect to the positions of their origin.
Reactions were prepared as follows. MmeI, SAM, and NEB Buffer 4 (10x) were obtained from New England Biolabs. 32 mM SAM was diluted 1:20 (-> 1.6 mM) in 1x NEB4 buffer.

- 173.9 ul DNA
- 664.5 ul dH20
- 100.0 ul NEB4(10x)
- 1.6 ul 1.6 mM SAM
- 60.0 ul MmeI (2U / ul)

Reaction was prepared on ice, and reagents were well-mixed prior to adding enzyme. Split to 8 tubes of 125 ul, and incubated on a thermal-cycler for 30 minutes at 37'C. Did not bother to heat-inactivate enzyme. Instead, went directly to P:C:P purification as described in above (step c), except using only 2 volumes of ethanol instead of 2.5 volumes (this may increase yield of smaller fragments?) Digested fragments were resupended in 80 ul of TE.

Diagnostic 6% PAGE gel of MmeI-digested RCA material



1 1 kb ladder (Invitrogen, 2 ul)
2 100 bp ladder (NEB, 2 ul)

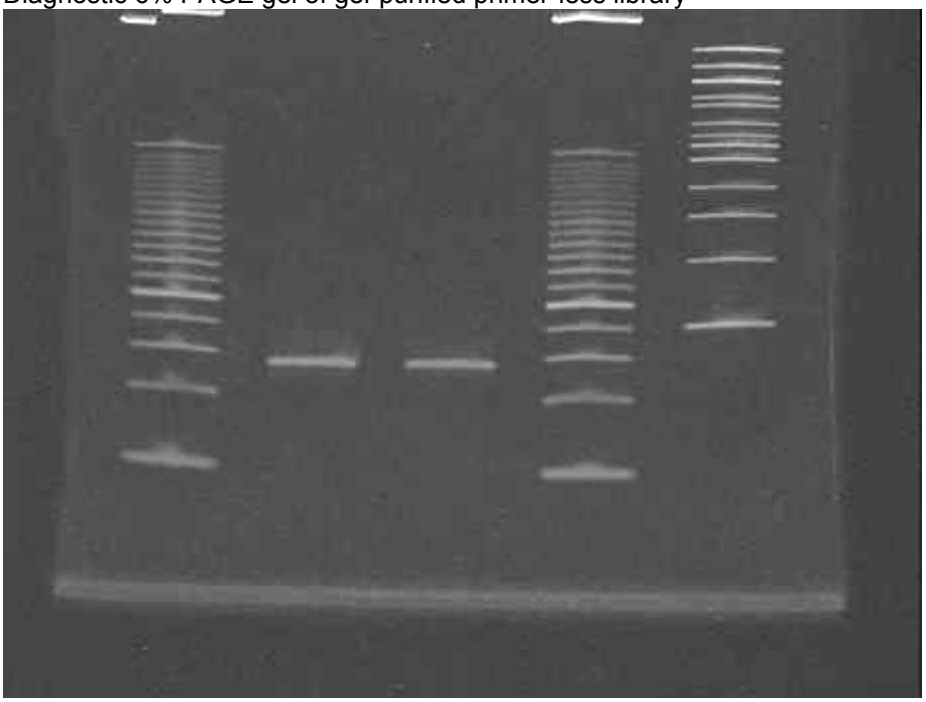
3 reference sample (2 ul)
4 library sample (2 ul)
5 25 bp ladder (Invitrogen, 2 ul)

We are expecting release of a band at ~70 bp in the library and reference lanes, which is what we see.

(h) PAGE purification of tag-T30-tag library

The full amount of each library was run on a 10-lane pre-cast 6% PAGE gel as above, using 4 lanes per library (20 ul of library + 5 ul of 5x dye) and leaving a blank spacer lane between the library-lanes of each type and any marker-lanes. A sharp band at approximately 70 bp was cut. Fragments from all lanes of each library type were combined, and the gel-extraction was carried out as described above (step d) except elution was for ~3 hours only. After P:C:P recovery, the DNA for each library was resuspended in ~20 ul of TE. A diagnostic gel was run to quantitate the recovered material.

Diagnostic 6% PAGE gel of gel-purified primer-less library



1 25 bp ladder (Invitrogen)
2 library (2 ul)
3 reference (2 ul)
4 25 bp ladder (Invitrogen)
5 100 bp ladder (NEB, 1 ul)

Estimated, based on relative intensities of bands of NEB 100 bp ladder, that the concentration of the library is ~12.5 ng/uL (and we have 18 ul remaining at this point)

(i) End-repair of tag-T30-tag library

The tag-T30-tag molecules contain 2 bp 3'-overhangs, consequent to MmeI digestion. Prior to ligating on primers, we need to repair these ends. Again, using the EpiCentre End-It DNA End Repair Kit that was used in step c).

Prepared reactions as follows:

- 8.50 ul of library fragments (12.5 ng/uL -> ~100 ng)
- 1.25 ul of 10x EndIt Buffer
- 1.25 ul of 10x EndIt ATP
- 1.25 ul of 10x EndIt dNTPs
- 0.25 ul of EndIt Enzyme

Total volume = 12.5 ul.

Incubated reaction at room-temperature for 45 minutes, then moved directly to 4'C. Increased volumes to 50 ul by adding 40 ul of TE, then P:C:P extracted as described above. In this case the precipitation step was allowed to go overnight at -70'C. Recovered DNA was resuspended in 8 ul of TE.

(j) Ligation of FDV2 (PR1-F) and RDV2 (PR1-R) primer oligonucleotides
The primer-adaptors (dsDNA, FDV2 and RDV2) were prepared by annealing fully complementary oligonucleotides (100 uM, HPLC-purified) by mixing 1:1 (final concentration of 50 uM), heating to 95'C for 10 minutes, and allowing to cool slowly over the course of an hour.
In "annealed" format, FDV2 and RDV2 are as follows:

- FDV2:
  5'-AACCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT -3'
  3'-TTGGTGATGCGGAGGCGAAAGGAGAGATACCCGTCAGCCACTA -5'
- RDV2:
  5'-AACTGCCCCGGGTTCCTCATTCTCT -3'
  3'-TTGACGGGGCCCAAGGAGTAAGAGA -5'

The FDV2 and RDV2 molecules are unphosphorylated, and therefore not expected to be able to self-self ligate nor to ligate to one another. The end-repaired ligated molecules are phosphorylated, and therefore an excess of FDV2 and RDV2 is used here to minimize concatamerization events for library molecules. Ligations of primer-adaptors to the library molecules were blunt-blunt and therefore conducted in the presence of PEG to improve ligation efficiency. Reaction was set up as follows:

- 12.3 ul of dH20
- 8.0 ul of purified, blunted library fragments (~100 ng -> ~2 pmols)
- 1.0 uL of RDV2 (50 uM -> 50 pmols)
- 1.0 uL of FDV2 (50 uM -> 50 pmols)
- 2.5 ul of 10x T4 Ligase Buffer (NEB)
- 21.2 ul of 40% PEG (40% polyethylene glycol 8000)
- 2.0 uL of T4 Ligase (NEB, 2000 U/uL)

Reaction was prepared at room-temperature by mixing all reagents except the PEG and ligase. Then PEG was added and mixed in, the ligase was added and mixed in. Reaction was put at 16'C overnight. To purify, increased reaction-volumes to 100 ul with TE and P:C:P purified as usual. Resuspended pellet in 10 ul of Buffer EB.

(k) PAGE size-selection of paired-tag library
Full amounts were run on a 10-well 6% PAGE gel along with the appropriate ladders. The gel was run far enough such that unligated RDV2 and FDV2 (present in great molar excess relative to the library) were expected to have run off the gel. We expected and observed a triplet of bands of the appropriate size (resulting from RDV2/library fragment/RDV2 ligation, RDV2/library fragment/FDV2 ligation, or FDV2/library-fragment/FDV2 ligation). The regions containing the full triplets were cut and gel-purified as described above in step (d), except elution was for 3 hours only, and ethanol precipitation was overnight at -70'C. Samples were each resuspended in 20 ul of Buffer EB.

(l) Nick translation to eliminate nicks in dsDNA library
As only the library molecules were 5'-phosphorylated in the ligation, the ligation products are expected to contain nicks that must be repaired. Moving forward with half of the remaining material, nick translation was performed as follows:

- 10.0 ul of library (assuming 100% recovery, this should be 50 ng of tag-T30-tag molecules plus the mass of ligated primer-adaptors)
- 0.5 ul of dNTP mix (25 mM, so final concentration of 500 uM for each nucleotide)
- 2.5 ul of 10x NEB-2
- uL of E.coli DNA polymerase I (NEB, 10 U/uL)
- 11.0 ul of dH20

Reaction was prepared and mixed on ice. Incubated at 16'C for 30 minutes. To purify, increased volume to 100 ul with TE and P:C:P purified as described above. Resuspended sample in 10 ul of TE.

(m) PCR amplification of paired-tag library

The reasons for PCR at this stage are first, to increase the amount of library material that we have to work with, and second, to eliminate extraneous ligation products in a single step. The only ligation products that should result from the PCR described in this step have tag-T30-tag flanked by properly oriented RDV2 and FDV2 on either side (note that the T30 segment itself is not symmetric, and therefore may be oriented in either direction relative to RDV2 and FDV2 segments).

As we are PCR-amplifying a complex mixture, it is critical to stop the reaction before primer molecules are exhausted. The reasons are first, that we observe that library molecules will begin to serve as primers for one another once the intended primers have run out, and second, that the library molecules contain enough similarity (~100 out of ~134 identical bases) such that after denaturing, it is unlikely that a given single-stranded library molecule will reanneal to its exactly complementary partner; the resultant library that has denatured and reannealed after primer has been exhausted may contain many "hybrid" library molecules.

PCR amplification was performed on a real-time PCR machine (MJ Research Opticon 2) as follows:

- Master mix:

	Per 50 ul	x20 (total volume of 1000 ul)
10x PCR Buffer	5	100
25 mM (each) dNTPs	0.4	8
50 mM MgCl2	1.5	30
Platinum Taq	0.2 ul	4
Water	42.7	853
RDV2-T (100 uM)	0.1	2
FDV2-T (100 uM)	0.1	2
SybrGreen (200x)	0.0025	0.5

- Master mix was split to two tubes of 499.5 ul, and 0.5 ul of library material was added. PCR was split to 8 tubes of 50 ul each (total volume of 400 ul) to run on thermal cycler.
- Thermal cycling was follows:
  - 1 94'C for 2 minutes
  - 2 94'C for 30 seconds
  - 3 55'C for 30 seconds
  - 4 72'C for 90 seconds
  - 5 Go to step 2

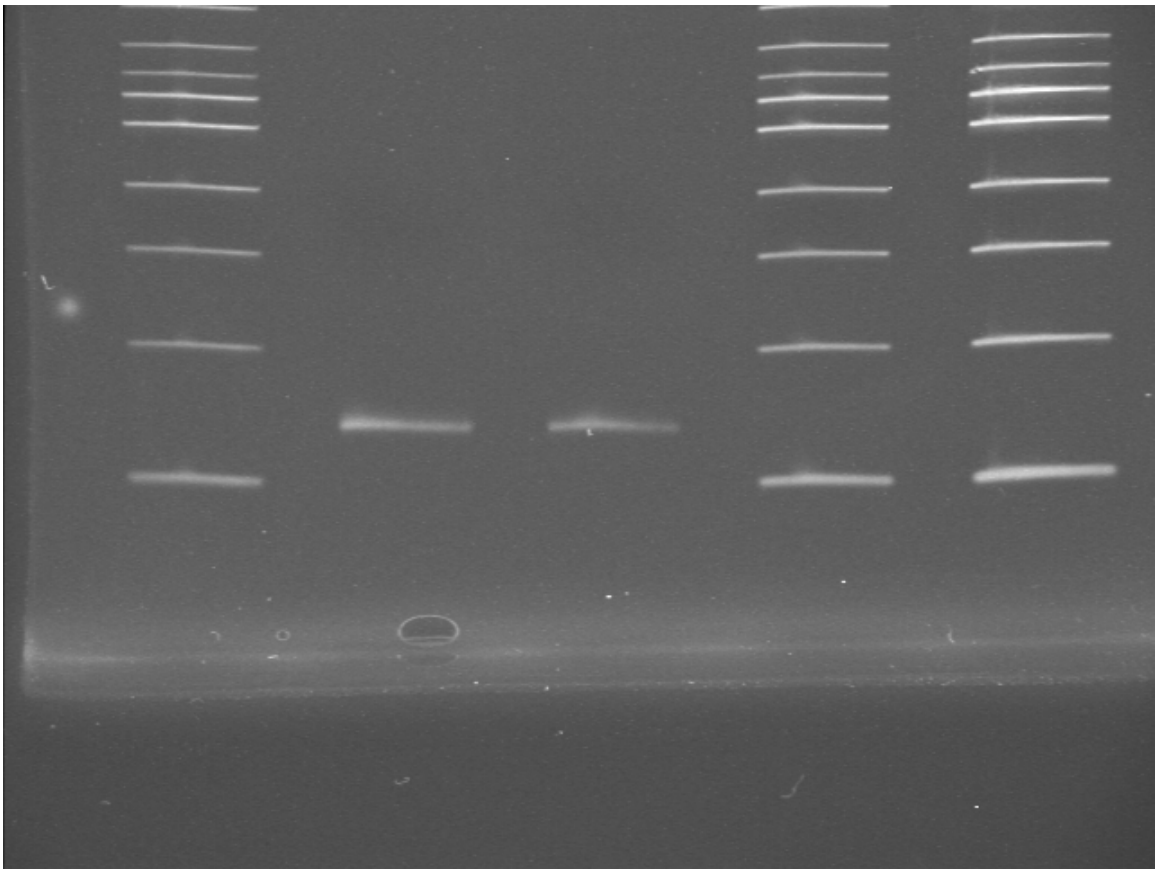
Reactions were stopped after 15 cycles because the quantity of DNA appeared to be beginning to plateau. Reactions from each library were combined to individual tubes and purified with Qiagen Qiaquick columns as per manufacturer's recommendations for PCR product purification. Resuspension was in 100 ul of Buffer EB. Subsequently decided this was too high a volume for the next step, so the samples were ethanol precipitated (as in P:C:P protocol, but no phenol-extraction was done), washed and resuspended in 10 ul of TE.

(n) PAGE size-selection of paired-tag library

Reactions were run on a 6% PAGE gel; the 'final' library bands, sharp bands at ~135 bp, were cut, eluted and purified as previously. This is the last purification step- it is therefore critical to try and get as tight of a gel-purification as possible to minimize contamination from any non-library molecules that might be present (in past iterations at developing this library construction protocol, these have generally been observed to occur from genomic E. coli fragments where MmeI sites coincidentally fall the necessary distance from one another to generate a fragment approximately equal in length to tag-T30-tag molecules). We suggest running the PAGE gel with NO ladders (as these molecules can also be frequent contaminents) and using a razor blade in a guillotine-type motion, rather than a scalpel.

After P:C:P with overnight precipitation, recovered library material was resuspended in 10 ul of TE. To quantify the library, a 6% PAGE gel with appropriate markers was run.

Diagnostic 6% PAGE gel of final libraries



1 0.125 ul of NEB 100 bp ladder
2 1.000 ul of library
3 1.000 ul of reference
4 0.250 ul of NEB 100 bp ladder
5 0.500 ul of NEB 100 bp ladder

Based on relative intensities of library and ladder bands, estimating concentration at 2 ng/uL, so we have ~9*2 = 18 ng of each library remaining. If the library is 135 bp, then the concentration is ~23 nM. The libraries were diluted in TE to various levels for use in emulsion PCR, and both the original libraries and their dilutions were stored at -20'C.

(o) Library validation via cloning and Sanger sequencing
To validate the expectation that library tags would be E.coli derived and paired, reference fragments were cloned with the Invitrogen TOPO-4 kit, PCR'd using M13F/M13R, and Sanger sequenced (single read per clone).

Although 96 PCR products were submitted for sequencing, 20 of these came back as either garbage reads, or vector- or contaminant related. The remaining 76 inserts appeared be appropriately flanked by the RDV2 and FDV2 segments, as expected.

Of these 76:

- One is a 6 bp insert (TTATCA)
- One is a E.coli genomic fragment (65 bp in length; 63/63 100% match to E.coli MG1655 genome on BLAST)
- One is a E.coli genomic fragment (70 bp in length; 69/69 100% match to E.coli MG1655 genome on BLAST)

- One contains the RDV2 primer flanked by ~27 bp with no significant matches in the NCBI database
- Seventy-two contain two tags separated by the T30 segment, as expected.

Of these 72, tag lengths had the following distribution:

- 1 tag was 9 bp
- 1 tag was 11 bp
- 2 tags were 13 bp
- 1 tag was 14 bp
- 1 tag was 15 bp
- 2 tag was 16 bp
- 73 tags were 17 bp
- 62 tags were 18 bp
- 1 tag was 22 bp

In terms of pairing, tags matched E.coli genome as follows:

- 4 were situations where one or neither tag had any perfect matches to the E. coli genome (most likely due to sequencing errors or non-canonical sequence)
- 1 was “unpaired” in that tags both matched unique locations but did not appear to originate from the same genomic regions
- 67 were matched the E.coli genome as paired tags (identically oriented with intertag distance falling within expected constraints)

For these 67 paired tags, the distance distribution of the paired tags was 951 +/- 90 bp. The minimum distance was 729 bp and the maximum distance was 1162 bp. Thus, the pairing rate for the 68 reads where we were able to map the origin of both tags is therefore 67/68 = ~98.5%.
A minimal estimate of the fraction of emulsion-PCR-amplifiable molecules in the library that represent paired E.coli tags with a T30 segment separating them is therefore 67/76 = 88%. The actual fraction may be slightly higher if the 4 reads where one or both tags were unmatchable actually do represent paired reads that were not matchable due to Sanger sequencing errors or differences between the “R” strain and the canonical genome sequence reference.

Although our sample size here is small (n=72), we noticed deviations from 25/25/25/25 frequencies in the tag sequences that may be significant trends and therefore list them here.
The numbers in the first column represent the tag base position relative to either its junction with one or the other primer or with the T30 segment. We chose to list these frequencies separately because the primer/tag ligation was blunt/blunt and the T30/tag ligation was based on a single-base (T/A) overhang. The strand from which the base frequencies were tabulated is such that these frequencies are what we wouldexpect to see if sequencing by extension from the primer or T30 segment (5’->3’). Numbers in parentheses are the actual counts (as opposed to frequencies).

PRIMER/TAG JUNCTION

	A	G	C	T
+1	0.315 (45)	0.315 (45)	0.154 (22)	0.217 (31)
+2	0.340 (49)	0.104 (15)	0.208 (30)	0.347 (50)
+3	0.292 (42)	0.208 (30)	0.146 (21)	0.354 (51)
+4	0.229 (33)	0.250 (36)	0.264 (38)	0.257 (37)
+5	0.333 (48)	0.194 (28)	0.243 (35)	0.229 (33)

T30/TAG JUNCTION

	A	G	C	T
+1	0.299(43)	0.194 (28)	0.326 (47)	0.181 (26)
+2	0.299(43)	0.319 (46)	0.146 (21)	0.236 (34)
+3	0.312(45)	0.222 (32)	0.222 (32)	0.243 (35)
+4	0.188(27)	0.236 (34)	0.264 (38)	0.312 (45)
+5	0.252(36)	0.196 (28)	0.273 (39)	0.280 (40)

SUPPLEMENTAL NOTE 2
Amplification Protocol

The ePCR, enrichment, and arraying protocols collectively take ~2 days.

Table 1. Optimizations of the Dressman protocol for stronger amplification

Parameter	*Dressman*	*Optimized*
Aqueous : oil ratio	1:2	1:6
Nucleotide concentration	1.0mM	3.5mM
$MgCl_2$ concentration	6.7mM	18.8mM
Taq polymerase	135U	270U
Extension time	30 seconds	75 seconds
PCR cycles	40	120

PROTOCOL

Bead-loading

- To 100ul 1um Dynal MyOne paramagnetic streptavidin beads from stock vial add 100ul Bind and Wash buffer (5mM Tris-HCl, 0.5mM EDTA, 1.0M NaCl, pH 7.5) in 1.5ml eppendorf. Mix and remove all liquid. Wash 2x in 200ul Bind and Wash buffer, and resuspend in 198ul (or 180ul, see below) Bind and Wash buffer.
- Add 2ul 1mM dual-biotin forward primer. Mix, and allow to stand for 20'. Periodically mix by pipetting.
- Remove all liquid, and wash 2x in 200ul Bind and Wash buffer, then 1x in 200ul TE. Remove liquid and resuspend in 200ul TE. These beads should now be at 5x10^9 beads/ml.

Emulsion PCR

- Prepare oil phase (make 6 tubes of the following):
    - 545ul light mineral oil (Sigma)
    - 450ul 10% Span 80 (Sigma) in light mineral oil
    - 4.0ul Tween 80 (Sigma)
    - 0.5ul Triton X-100 (Sigma)
    - -10% Span 80 solution reduces pipette errors caused by the high viscosity of Span 80. Mix 10ml at a time using syringes for accurate measurement.
    - -use reverse pipetting (or positive displacement) with Tween and Triton for accurate delivery
    - -after all ingredients are dispensed, vortex to mix thoroughly
- Prepare 960 ul aqueous phase:
    - 1x MgCl2- PCR buffer (Invitrogen),
    - 18.8mM MgCl2 (Invitrogen),
    - 3.5mM (each) dNTP mix (Invitrogen),
    - 25uM reverse primer (IDT),
    - 50nM unmodified forward primer (IDT),
    - 60ul forward primer-loaded beads (Dynal),
    - 270U Platinum Taq (Invitrogen), 1.00ul template DNA
- Put 4 2ml Corning cryogenic vials in center of VWR 565 closed-loop magnetic stirrer set to 1400rpm. Place flea-size stir bar in each vial, add 400ul oil phase, and then add 75ul aqueous phase, dropwise over 1 minute to each vial. Stir for 30 minutes. Pipette 50ul into each of 8 200ul PCR tubes (total of 32 tubes).
- PCR using the following program:
    - 1 94C 2 min
    - 2 94C 15 sec

    - 3 57C 30 sec
    - 4 70C 75 sec
    - 5 goto 2, 119 more times
    - 6 72C 2 min
    - 7 4C forever
- Pool contents of each 8 200ul tubes in a 1.5ml tube and break emulsion as follows:
    - add 800ul NX2 buffer (100mM NaCl, 10mM Tris-HCl pH 7.5, 1mM EDTA, 0.1% Triton X-100), vortex 20 seconds, spin 1.5' at 13200rpm.
    - remove most liquid, add 800ul NX2, vortex 20 seconds, spin 1.5' at 9000rpm
    - remove most liquid, add 700ul NX2, vortex 20 seconds, spin 1.5' at 9000rpm
    - remove most liquid, add 600ul NX, vortex 20 seconds, spin 1.5' at 9000rpm
    - remove most liquid, then place on magnet and remove all liquid
- resuspend in 20ul TE and transfer to a new tube, pooling all 12 tubes into 1
- wash 2x in 250ul TE; remove all liquid with magnet
- Denature free reverse strands by adding 250ul 0.1M NaOH to tube and incubating at RT for 10'; wash 1x in 0.1M NaOH, and 3x in TE; resuspend in 60ul TE

The following primers were used for emulsion PCR:

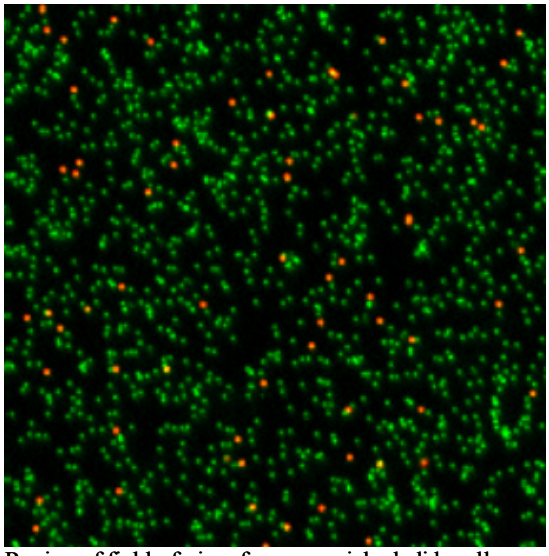
- Dual-biotin forward primer: 5'-DualBio/CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT-3'
- Unmodified forward primer: 5'- CCTCTCTATGGGCAGTCGGTGAT-3'
- Reverse primer: 5'-CTGCCCCGGGTTCCTCATTCTCT-3'

SUPPLEMENTAL NOTE 3

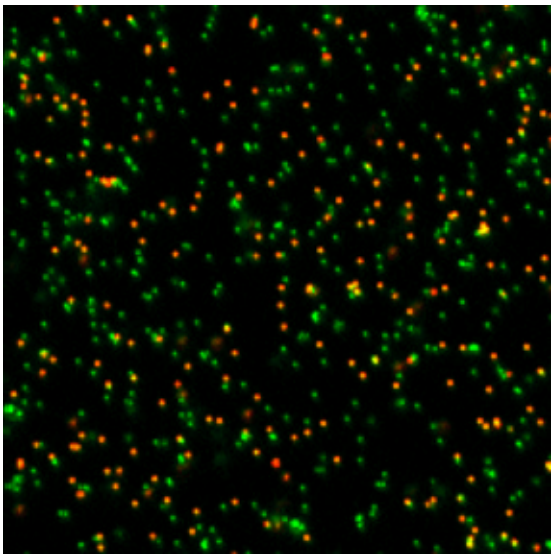
Bead Enrichment Protocol

Enrichment of amplified beads is accomplished by hybridization to larger, less dense nonmagnetic 'capture' beads bearing a DNA sequence complementary to ePCR amplicon sequence. Once hybridized, this density difference between the amplified bead:capture bead complexes and un-amplified beads is exploited by centrifugation through glycerol (60% v/v, 1 minute at 13200rpm). The less-dense complexes remain in the supernatant, while the un-amplified beads form a pellet. Supernatant is recovered, complexes are melted with NaOH, and enriched amplified beads are recovered by magnetic separation.

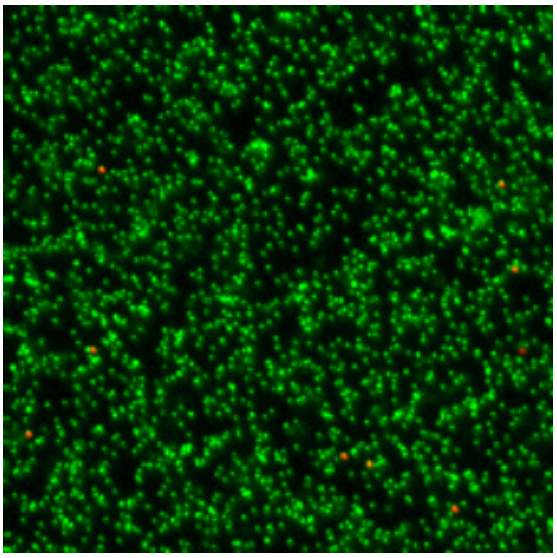
A proof-of-concept experiment was performed with a population of beads where approximately 8% were 'amplified', in that they had the sequence of interest. Following enrichment, 42.5% of beads in the supernatant fraction were found to bear the sequence of interest on the basis of hybridization of a fluorescent probe (10uM probe in 6x SSPE w/ 0.01% Triton X-100 at 56C for 5'), and 0.7% of beads in the pellet fraction were 'amplified', for an enrichment of ~5.3x, with a recovery of ~20%.



Region of field of view from unenriched slide; all beads are green, template beads are red (200x200 pixels from 1200x1600 pixels, total of 39,79 beads)



Region (200x200 pixels) from supernatant; 5.2x enrichment



Region (200x200 pixels) from pellet slide

The following oligonucleotides were used for enrichment:

Biotinylated capture oligo:

- 5'Bio/CGTACCCCGCTTGGTCTTTCTCCCGTACCCCGCTTGGTCTTTCTCCCTGCCCCGGGTTCCTCATTCTCT-3'
- Sequence of interest: 5'Bio/CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATAGAGTGGTGGACGACAGCTCTCACATAGAGAATGAGGAACCCGGGGCAG-3'
- Cy3 probe: 5'-Cy3/CTGCCCCGGGTTCCTCATTCTCT-3'

SUPPLEMENTAL NOTE 4

Bead Arraying Protocol

Beads are poured in a 5% acrylamide gel onto a Bind Silane-treated 40mm round #1.5 glass coverslip (Bioptechs). The gel geometry is formed using a teflon-coated glass microscope slide as a template (Erie Scientific). A slide with a round 14mm well is thus used to create a circular gel approximately 30 microns thick. Polymerization is slowed by using reduced amounts of catalyst such that the beads settle into a single focal plane at the surface of the gel (the coverslip is inverted to that the exposed gel surface is facing down).

Coverslips are treated as follows:
Wash 20' in 1% Triton X-100
During washing, prepare Bind Silane treatment:

- 73 ul acetic acid
- 1300uL Bind Silane (Amersham)
- 350mL dH2O

Stir for 15' with magnetic stir bar
Rinse coverslips in dH2O
Incubate for 1 hour in Bind Silane solution with gentle shaking
Wash 3x in dH2O, and 1x in 95% EtOH
Allow to dry and store in vaccum dessicator

To create a bead array, mix the following:

- 60 ul beads from ePCR, resuspended in 6.5ul TE
- 1.25ul 40% acrylamide/bis (19/1, Roche)
- 0.5ul Rinohide gel strengthener (Molecular Probes)
- 0.5ul 5% TEMED
- 0.75ul 0.5% ammonium persulfate

Pipette 9.5ul mixture onto teflon-coated slide, then drop coverslip on top and invert.
Polymerize for 1 hour at 25C.
Remove coverslip (with gel attached) and immerse in dH2O to wash out un-polymerized acrylamide and loose beads. Assemble into flowcell immediately.

SUPPLEMENTAL NOTE 5

Instrument Components and Cost

The sequencing instrument is composed of two main functional assemblies: an automated fluorescence microscope, and a computer-controlled flowcell / fluidics system. The automated microscope has the following main components:

- Nikon TE2000-E inverted epiflourescence microscope
  - Automated filter turret
  - LWD collector lens (NA 0.50) for brightfield illumination
  - Automated focus with linear encoder (50nm resolution)
  - 20x Plan Apo DM (phase) objective
- Prior XY stage with 1mm screw pitch and linear encoders
- Sutter Lambda LS xenon light source (175W)
- Sutter Lambda 10-3 filter wheels (excitation and emission)
- Sutter SmartShutter for light path
- Hamamatsu electron-multiplied camera
  - 1000 x 1000 pixels
  - 14 bit dynamic range
  - 30 fps readout with no binning

The fluidics system is composed of the following main parts:

- Bioptechs flowcell
  - 30mm viewable aperture
  - resistive heating from 25 – 60C
  - 24 x 14 x 0.25 mm reagent chamber
  - all-glass chamber for Koehler illumination
- Alcott autosampler
  - 96 addressable reagent wells
  - refrigerated stage
- Superlogics A/D modules for temperature control and washing

A detailed parts list:

NIKON

1	MEA51001	TE2000-E2 Inverted Microscope	$ 13,295.25	$ 13,295.25
1	MEF55010	T-HUBC hub controller	$ 3,752.50	$ 3,752.50
1	MEF55000	T-RP remote control pad	$ 1,805.00	$ 1,805.00
1	MPF52061	Univ power supply 110-240V	$ 180.50	$ 180.50
1	MAE15001	Lamphouse HMX2	$ 502.55	$ 502.55
1	MEE59900	T-DH dia-illuminator 100W	$ 1,562.75	$ 1,562.75
1	MAE15002	Collector lens for QH100	$ 239.40	$ 239.40
1	MBF13240	Lamp socket F/HMX 100W QI	$ 244.15	$ 244.15
1	MXA20425	FXA-HMX2 adpt f/halogen	$ 524.40	$ 524.40
1	84125	Microphot-FXA L.L. 12V-100W BU	$ 26.60	$ 26.60
1	91141-IN	T-BP R100 optical path prism	$ 608.00	$ 608.00
1	MEF42252	TE2-PS 100W pwr supply	$ 741.95	$ 741.95
2	79035	Power cord	$ 11.40	$ 22.80
1	91155	Null modem cable	$ 11.40	$ 11.40
1	MEP51300	T-N6 sextuple nosepiece	$ 453.15	$ 453.15
1	MBN11710	Filter 45mm NCB11	$ 67.45	$ 67.45
1	MEL3000	Diaphot condenser turret	$ 506.35	$ 506.35
1	MEL36200	Diaphot condenser Lwd lens	$ 579.50	$ 579.50
1	MEH41200	Te-C Lwd Ph2 module	$ 93.10	$ 93.10
1	MEV51100	T-FLMC motorized cassette holder	$ 1,724.25	$ 1,724.25
1	MRD30200	Cfi plan apo dm 20x objective, NA 0.75	$ 2,782.55	$ 2,782.55
1	63-561	Tmc antivibration table, 30"x48"	$ 2,875.00	$ 2,875.00

1	81-301-02	front support bar 30"x48"	$ 160.00	$ 160.00
				$ 32,758.60

PRIOR

1	PR 5H152V2	Joystick control for XY-Axis stage controller	$ 600.00
1	PRH29XYE	ProScan XY controller w/ encoder	$ 4,775.00
1	PR5H117E2TE	Stage, Invtd, Flat Top with scales for TE2000	$ 11,800.00
			$ 17,175.00

SUTTER

1	SULBLSOF17	Lambda LS Lamphouse 175 Xenon (ozone free)	$ 4,450.00
1	SULGN27	Adapter for Liquid Light Guide to Nikon TE2000	$ 700.00
1	SUO661176	175W Xenon bulb (ozone free)	$ 631.00
1	SULLG	Liquid Light Guide for Lambda LS xenon lamp house	$ 1,500.00
1	SULB10-3	Lambda 10-3 controller unit	$ 2,400.00
1	SULB10-NWIQ	10 position 25mm filter wheel with Smart shutter	$ 3,450.00
1	SULB10W	2nd 25mm Filter Wheel - no shutter	$ 2,500.00
1	SU10N27EM	FWheel emission adapter f/ Nikon TE2000	$ 350.00
1	SU10N27EC	FWheel excitation adapter f/ Nikon TE2000	$ 300.00
1	IQ35-N27	Smart Shutter for HMX lamphouse	$ 1,100.00
			$ 17,381.00

HAMAMATSU

1	HAC9100-02	EM CCD Camera 1kX1k CCD W/CBL & Pwr Sply	**$ 34,500.00**

UNIVERSAL IMAGING

1	40003	Metamorph Premier Software, current user	$ 9,100.00
1	40095	Hamamatsu DCAM driver	$ 1,600.00
			$ 10,700.00

CHROMA

1	86017	FITC / Texas Red, single emitters & exciters, polychroic	$ 1,025.00
1	86022	Cy3 / Cy5, single emitters & exciters, polychroic	$ 1,025.00
			$ 2,050.00

ALCOTT CHROMATOGRAPHY

1	719/00000/01R	719 ALR autosampler, no valve assembly	$ 13,995.00
1	720/25815/00	Titer plate adaptor	$ 70.00
			$ 14,065.00

BIOPTECHS

1	060319-2-0303	Temperature controller, no alarm	$ 2,000.00	$ 2,000.00
1	060319-2-03	FCS2 chamber only, 30mm aperture	$ 1,300.00	$ 1,300.00
1	060319-2-1242	low dead volume top, drilled	$ 510.00	$ 510.00
1	060319-2-0049	FCS2 connector ass'y, mod for slide only	$ 495.00	$ 495.00
1	130119-3153	microaqueduct machine changeover	$ 300.00	$ 300.00
5	130119-5-HG	specialty microaqueduct slide	$ 70.00	$ 350.00
20	1916	specialty gasket for 20mmx20mm square	$ 5.00	$ 100.00
1	40-1313-0319	40mm round coverslips, 250/pk	$ 200.00	$ 200.00

					$ 5,255.00
TECAN SYSTEMS					
1	726802	XL3000 RS232 1/4-28 4-port syringe pump	$	924.00	
1	725030	syringe, 500ul reagent	$	75.00	
			$	**999.00**	
SUPERLOGICS					
1	8017	8 Channel Analog Input, 16-Bit, Data Acquisition Module	$	195.00	
1	8024	4 Channel 14-Bit Analog Output Module (RS-485)	$	269.00	
1	8060	4 TTL Digital Input/4 Relay Output Module (RS-485)	$	115.00	
1	8520	RS-232 to RS-485 Converter Module	$	94.00	
			$	**673.00**	
		TOTAL COST	$		135,556.60

SUPPLEMENTAL NOTE 6

Sequencing Protocol

Sequencing by degenerate ligation is a process where the following steps are executed cyclically to interrogate each base of the template sequentially:

- Hybridize 'anchor primer' complementary to common library sequence
- Ligate pool of fluorescently-labeled 'query primers' specific to one tag-position
- Image to determine which primer pool ligated to each bead
- Strip anchor::query primer complex
- Repeat

Anchor primers used had the following sequences (U = deoxyuridine):

- T30UIA 5'-GGGCCGUACGUCCAACT-3'
- T30UIB 5'-CGCCUUGGCCUCCGACT-3'
- PR1UI0N 5'-CCCGGGUUCCUCAUUCUCT-3'
- LIGFIXDD 5'-Phos/AUCACCGACUGCCCA-3'
- LIGFIXD2T30A 5'-Phos/AGUUGGAGGUACGGC-3'
- LIGFIXD2T30B 5'-Phos/AGUCGGAGGCCAAGC-3'

Query primers used were nonamers which were degenerate at all positions excepy the query position. At the query position, only one base was present for a given fluorophore. For example, the pool of probes used to query position five was composed of the following four label-subpools:

- Cy54NA 5'-Phos/NNNNANNNN/Cy5--3'
- Cy34NG 5'-Phos/NNNNGNNNN/Cy3-3'
- TexasRed4NC 5'-Phos/NNNNCNNNN/TR-3'
- FRET4NT 5'-Phos/NNNNTNNNN/FRET-3'

Anchor primers were hybridized in the flowcell (100uM primer in 6x SSPE) for 5' at 56C, then cooled to 42C and held for 2'. Excess primer was then washed out at room temperature with Wash 1E (10mM Tris-HCl pH 7.5, 50mM KCl, 2mM EDTA pH 8.0, 0.01% Triton X-100) for 2'.

Query primers were ligated in the flowcell (8uM query primer mix (2uM each subpool), 6000U T4 DNA ligase (NEB), 1x T4 DNA ligase buffer (NEB)) at 35C and held for 30'. At the end of the reaction, excess query primer was washed out at room temperature with Wash 1E for 5'.

Imaging was performed as described in Supplemental Note 8.

Anchor::query primer complex was stripped with USER (NEB), a combination of uracil DNA glycosylase and endonuclease VIII. To perform the stripping reaction, the following protocol was executed in the flowcell:

- Incubate 150uL stripping mix (3 ul USER (NEB), 150 ul TE) for 5' at 37C
- Raise temperature to 56C and hold 1'
- Wash for 1' with Wash 1E; temperature gradually decreases
- Incubate 150 ul fresh stripping mix for 5' at 37C
- Wash for 5' with Wash 1E; temperature gradually decreases

SUPPLEMENTAL NOTE 7

Image Processing and Basecalling Algorithms

The sequencing system software can be divided into four subsets:

- protocol automation control
- image acquisition
- data extraction
- base-calling and read-mapping

Image acquisition is performed mainly by the commercial imaging package Metamorph and makes use of custom journals which interface with a compiled Matlab routine. Protocol automation control is achieved with a set of Matlab functions utilizing the Instrument Control Toolbox. Data extraction is performed by a series of binaries written in C/C++, and base-calling and read-mapping is performed by a series of Matlab and Perl scripts. This note will first outline the process of generating data from beginning to end, then step through each phase in more detail.

Sequencing Process

For each base in the tag to be queried (26 total), we perform a hybridization reaction, a ligation reaction, a cycle of imaging, and a primer stripping reaction. The Matlab code which runs the fluidics system (i.e. everything not associated with image acquisition) consists of a series of functions, "wobbler_hyb", "wobbler_react", and "wobbler_strip" which perform each task. The Metamorph journals which perform the image acquisition coordinate their activity with that of the Matlab functions. Thus, Metamorph opens a Matlab instance to execute the proper sequence of protocol commands (strip, hyb, react) and then commences imaging upon a normal return.

To determine the identity of each bead in the array at a given position, after performing the biochemical sequencing reactions, each field of view ("frame") is imaged with four different wavelengths corresponding to the four fluorescent nonamers used. All images from a cycle are saved with sequential filenames in a cycle directory, where the number of images is 4 x the number of frames (usually several thousand images). Cycle image data are saved into a directory structure organized for downstream processing.

Data extraction requires three types of image data: 1) brightfield images to demarcate the positions of all beads in the array; 2) fluorescence images acquired with a labeled primer complementary to a common region of the library such that all amplified beads will fluoresce, allowing identification of amplified beads; 3) sets of fluorescence images acquired during each sequencing cycle. The data extraction software identifies all objects in the brightfield images, then for each such object, computes an average fluorescence value for each sequencing cycle. A list is generated for each cycle containing such an average for each bead in the array.

Protocol Automation Control

To automate the sequencing protocol, we need to control the following pieces of hardware:

- Alcott 96-well autosampler
- Cavro syringe pump
- Lee 3-way valve
- Bioptechs flowcell temperature controller

The valve and temperature controller are both controlled by a set of Superlogics serial analog/digital I/O modules. We thus have written software to control 3 pieces of hardware; all devices communicate with the computer via RS232 instruction sets. We have written a 'driver' function for each piece of hardware which issues an RS232 command to the appropriate device in the format expected, and waits for a response, if any, before returning. The full list of such functions is below:

- autosampler.m
- bioptechs.m
- cavro.m
- relay.m

Sitting on top of this hardware interface layer is a set of functions corresponding to 'basic commands' one would need to do something useful with the hardware. Examples of such commands would be to draw a reagent from a well (which requires movement of the autosampler and syringe pump as well as setting of the temperature controller and opening of the needle valve) or pass wash buffer through the flowcell (which requires movement of the syringe and setting of the temperature controller). The full list of such functions is below:

- ll_draw_reagent.m
- syringe.m
- set_slide_temp.m
- get_slide_temp.m
- needle_open.m
- needle_close.m
- microaqueduct.m
- ll_needle_wash.m
- syringe_initialize.m

The final layer of abstraction is the set of functions which integrate basic commands into useful protocol steps, of which there are three: primer hybridization, primer stripping, and general enzymatic reactions (which is used for query nonamer ligation). Each of these functions is parameterized such that the user can specify which reagents to use, the temperature for the reaction, the duration of time to allow the reaction to proceed, etc. The functions are:

- wobbler_hyb.m
- wobbler_strip.m
- wobbler_react.m

Image Acquisition

To acquire the images from a single sequencing cycle, we raster across the array and acquire four images at each position, one each for each of the fluorophores present. Since we want this acquisition process to be fast, we start a sequencing experiment by generating a topographic map of the array by autofocusing at each position. We have found that this 'focal map' remains accurate during the course of the two-day, 24-cycle experiment. During the acquisition cycles, we then only need to visit each position in the map and acquire the images.

To generate the focal map, we first autofocus using a custom-written autofocus routine under brightfield illumination. We then refine this map by autofocusing using Metamorph's standard autofocus routine under fluorescence illumination after hybridizing a fluorescently-labeled primer to the bead array. We have found this is necessary when acquiring images with a plan apo objective because of its short depth-of-field. This dual-autofocus operation is not required when acquiring images with the long working distance plan fluor objective; the initial brightfield focal map is sufficiently accurate.

During acquisition, we use the "Save using sequential filenames" option in Metamorph to save all files into a single directory. We then execute Perl script "split_by_color.pl" to split this group of files into four directories, one for each of the four wavelengths, and re-number the files so that names correspond from one directory to the next.

Data Extraction

Data extraction is the process of generating a list of fluorescence values for each bead for each acquisition cycle. Thus, in a typical sequencing run of 104 cycles (for 26 bases of sequence per tag), each bead will have 104 values. To extract such data from the fluorescence images, we start each sequencing run by acquiring a series of brightfield images. Under brightfield illumination with a phase-contrast objective (we use a Nikon plan apo 20x DM, NA 0.75) and Abbe condenser (LWD, N.A. 0.50), beads appear as distinct objects having white centers and separated by dark rings. We thus are able to easily threshold such images and 'separate' beads which are very close together by then identifying all objects (where an object is defined as a set of 4-connected pixels).

To identify objects in all brightfield images, we execute the find_objects binary, which is essentially a C implementation of the Matlab bwlabel.m function. Find_objects takes a brightfield image and a threshold as input, and outputs an 'object' pseudo-image where each set of connected pixels in the input image is labeled with a unique object_id in the object image.

Before using these object images to extract fluorescence data from the sequencing cycle images, we must put them into register with the initial brightfield images such that the positions of beads in the brightfield correspond exactly to the positions of those beads in the fluorescence images. Register takes the brightfield image and one or more cycle images, and returns a text list of offsets to translate the cycle images by to put them in register with the brightfield images.

Once this set of object images is generated, it is used to define the pixel locations in each set of cycle images corresponding to beads. A mean is then computed for the fluorescence value for each bead, and a list is generated of all beads for a given cycle.

An identical process is carried out for set of images, such that we end up with a file containing, for each bead, a mean fluorescence intensity at each image-set, along with basic annotation information (raster position, X location, Y location, size in pixels).

For any given cycle, there are four data-points, corresponding to the four images taken at different wave-lengths to query whether that base is an A, G, C or T. The data for each channel is normalized such that it sums to a constant value, and then data from each bead is normalized to a four-dimension unit vector. Positive-valued four-dimensional unit-vectors can be projected into tetrahedral space, such that each vertex (e.g. (1,0,0,0)) is represented by one of the vertexes of the tetrahedron, such as that shown in Fig. 2c. Base calls for each bead at each cycle are assigned based on the maximum value of this normalized unit vector. The preponderance of each set of base-calls are observed to form a cluster. For each type of base-call at a given cycle, a median location within unit-vector space is calculated. Each base-call is assigned a quality score based on its Euclidean distance to the median unit-vector for the base to which it is assigned.

These raw base-calls (and accompanying quality metrics) are consolidated, yielding a discontinuous sequencing read for each bead. The next task is to match these sequencing reads against the reference genome. For each 17 to 18 bp tag, we are obtaining 6 bases in one direction and 7 bases in the other direction. This leaves an indeterminate gap of 4 to 5 bases central to each tag. We have a priori expectation (Fig. 3a), that the distance between the proximal and distal tags should be approximately 700-1200 bp. The read matching algorithm proceeds in the following order:

- Mask the two lowest quality base-pairs in each read (leaving 24 or 26)
- If a given tag has a single exact match to the reference sequence that meets the above distance constraints, assign it to that location.
- If a given tag has multiple exact matches to the reference sequence that meet the above distance constrains, discard it.
- Allowing for a single substitution, if a given tag has a single match that meets the above distance constraints, assign it to that location.
- Allowing for a single substitution, if a given tag has a multiple matches that meets the above distance constraints, discard it.

This yields a set of "final mappings" of sequencing reads to the reference genome. Estimated error rates for each bin are calculated based on differences between the reference genome and the bases called for

a given position. If all raw bases covering a given base are consistent, that base is called as such. If raw bases covering a given base are inconsistent, the estimated error rates for individual raw calls are taken into account, and the base is assigned to the call with the highest likelihood given the set of observed data. High confidence calls exceed log-likelihood threshold. A more stringent threshold is applied for positions with inconsistent data.

SUPPLEMENTAL NOTE 8

Protocol Cost Analysis

Note: We have included the cost of library construction labor and reagents (but not the cost of Sanger sequencing 96 clones) in the below calculation as if the library was constructed solely for this instrument-run. However, we note that there is no fixed relationship between the number of libraries made and the number of bases sequenced, as is the case for ePCR and sequencing (e.g. the same library could be sequenced over many experiments to greater depth). Not including library construction costs, our estimated costs drop to 8 cents per kilobase.

More Excel spreadsheets can be found at:
http://speedy.med.harvard.edu/PolonySeq/costs/

Cost Summary (per instrument run)

	Cost	N	Total
Emulsion Reagents (per plate)	217.92	1.00	217.92
Sequencing Reagents (per run)	774.60	1.00	774.60
Library Reagents (per run)	173.05	1.00	173.05
Labor (per run)	700.00	1.00	700.00
Equipment (per run)	254.47	1.00	254.47
Overhead (75%)			1305.90
Total			**3425.93**
Bases Per Bead			26
Bases per Run			30,100,000
Bases per Dollar			**8786**
Cents per Kilobase			**11.38**
Factor Cheaper Than Sanger			**9**

Cost Calculation for Sequencing Protocol Reagents (per experiment)

Enzymes (E)	**$555.89**	**71.8%**	0.7147
Oligos (O)	**$217.35**	**28.1%**	
Other (M)	**$1.36**	**0.2%**	
REAGENT COST	**$774.60**	**100.0%**	

Reagent	Step	uL (or mg) per step	# of uses	Total used (uL or mg)	Vendor	$ per tube	uL or mg per tube	cost per gel ($)	Category
ddNTPs (5 mM each)	TdT treatment	4	1	4	Amersham	136.00	400	1.36	M
5x tailing buffer	TdT treatment	20	1	20	Invitrogen	--	--	--	--
rTdT enzyme	TdT treatment	2	1	2	Invitrogen	276.00	100	5.52	E
USER enzyme	Primer stripping	6	26	156	NEB	252.00	250	157.25	E
10x T4 Ligase Buffer	Sequencing	15	26	390	NEB	--	--	--	--
T4 DNA Ligase (2000 U/uL)	Sequencing	3	26	78	NEB	252.00	50	393.12	E
						252.00			
Cy5 nonamers	Sequencing	3	26	78	IDT	240.30	460	40.75	O
Cy3 nonamers	Sequencing	3	26	78	IDT	240.30	408	45.94	O
CAL610 nonamers	Sequencing	3	26	78	IDT	307.80	1221	19.66	O
FRET nonamers	Sequencing	3	26	78	IDT	500.00	500	78.00	O
Cy3.T30.UI.A	Sequencing	1	2	2	IDT	150.00	100	3.00	O
Cy3.T30.UI.B	Sequencing	1	2	2	IDT	150.00	100	3.00	O
PR1.UI (1 mM)	Sequencing D(+1)..D(+9)	1.5	9	13.5	IDT	50.00	100	6.75	O
LigFixDD (1 mM)	Sequencing P(-1)..P(-9)	1.5	9	13.5	IDT	50.00	100	6.75	O
T30.UI.A (1 mM)	Sequencing D(-1)..D(-9)	0.75	9	6.75	IDT	50.00	100	3.38	O
T30.UI.B (1 mM)	Sequencing D(-1)..D(-9)	0.75	9	6.75	IDT	50.00	100	3.38	O
T30.LigFix.A (1 mM)	Sequencing P(+1)..P(+9)	0.75	9	6.75	IDT	50.00	100	3.38	O
T30.LigFix.B (1 mM)	Sequencing P(+1)..P(+9)	0.75	9	6.75	IDT	50.00	100	3.38	O

Cost Calculation for Emulsion PCR Reagents (per 12 strips or 96 tubes)

Enzymes (E)	$159.30	73.10%
Oligos (O)	$14.42	6.62%
Other (M)	$44.20	20.28%
REAGENT COST	$217.92	100.00%

Reagent	Step	uL per step	# of uses	Total used (uL)	Vendor	$ per tube	uL per tube	cost per gel ($)	Category
Platinum Taq	Emulsion PCR	4.5	12	54	Invitrogen	2950.00	1000	159.30	E
Stir Bar	Emulsion PCR	1	12	12	VWR	2.00	1	24.00	M
Nucleotides	Emulsion PCR	11.25	12	135	Invitrogen	1163.00	10000	15.70	M
MyOne Beads	Emulsion PCR	5	12	60	Dynal	300.00	4000	4.50	M
Dual-Biotin Forward Primer	Emulsion PCR	0.5	12	6	IDT	292.80	406	4.33	O
Free Reverse Primer	Emulsion PCR	4	12	48	IDT	39.10	186	10.09	O

SUPPLEMENTAL NOTE 9

Code

All code used for acquisition, protocol control, data extraction, basecalling, and read-mapping can be found at http://speedy.med.harvard.edu/PolonySeq/code/