Improved Protocols for the Illumina Genome Analyzer Sequencing System

Michael A. Quail,[1] Harold Swerdlow,[1] and Daniel J. Turner[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom

ABSTRACT

In this unit, we describe a set of improvements we have made to the standard Illumina Genome Analyzer protocols to make the sequencing process more reliable in a high-throughput environment, reduce amplification bias, narrow the distribution of insert sizes, and reliably obtain high yields of data. *Curr. Protoc. Hum. Genet.* 62:18.2.1-18.2.27. © 2009 by John Wiley & Sons, Inc.

Keywords: Illumina • Next-Generation • sequencer • protocols • Genome Analyzer

INTRODUCTION

Knowledge of the DNA sequence of an organism is the key to understanding how that organism exists. With it, we can define characteristics of genomes, and delineate differences between them, which, in turn, help us to understand genotype/phenotype relationships (Bentley et al., 2008; Mardis, 2008).

In the mid 1970s, several methods of sequencing DNA appeared around the same time (e.g., Sanger and Coulson, 1975; Maxam and Gilbert, 1977), but it was dideoxy DNA sequencing (Sanger et al., 1977) that proved to be the most versatile and practical approach. Over the following decades, dideoxy sequencing continued to be developed, and thirty years later, it is still used widely as the standard sequencing technology in many laboratories. The drawback of the method is that its throughput is limited, as sequencing is performed on single isolated templates, which means that large-scale sequencing projects are expensive and laborious, requiring ligation of target DNAs into cloning vectors, and amplification in *Escherichia coli*. Consequently, the human genome sequence (International Human Genome Sequencing Consortium, 2004), which was generated entirely by capillary sequencing using dideoxy chemistry, took hundreds of sequencing machines several years and the final sequencing phase cost ~300 million US dollars.

In 2005, the first of the next generation DNA sequencers, 454's GS20 (now Roche 454), became available commercially (Margulies et al., 2005), revolutionizing the paradigm of DNA sequencing. Instead of a single sequencing reaction generating a single sequence, the 454 introduced massively parallel sequencing, albeit on a relatively modest scale. The Roche 454 uses emulsion PCR to generate beads coated in amplicons derived from single template molecules. Hundreds of thousands of these beads are then sequenced in parallel, by pyrosequencing (Ronaghi et al., 1998). Images of the beads are analyzed to generate high-quality sequences. In this way, throughput is increased, cost is reduced, and cloning is avoided. The GS20 was capable of generating 20 megabases of sequence data per run, compared to <100 kb for a 96-well capillary machine, and the output has increased to several hundred Mb of 400 to 500 base reads per run with the Titanium version of this sequencing platform. A single Roche 454 sequencing run can thus generate sufficient data for many projects.

Current Protocols in Human Genetics 18.2.1-18.2.27, July 2009
Published online July 2009 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471142905.hg1802s62
Copyright © 2009 John Wiley & Sons, Inc.

QIAGEN Exhibit 1033.0001

However, although the long sequencing reads produced by the Roche 454 lend themselves particularly well to de novo assembly of bacterial genomes, there are sequencing applications for which read length is of lesser importance than total output and cost per gigabase. For example, within-species genetic variation can be identified by mapping sequence reads to a reference sequence and identifying positions that differ. The availability of high-quality reference sequences for many organisms allows such resequencing experiments to be performed with short-read sequences.

For the same amount of sequence data, short reads (25 to 100 bases) can currently be generated far less expensively than longer sequence reads. The two leading short-read Next-Generation sequencing platforms currently available are Applied Biosystems' SOLiD and the Illumina Genome Analyzer, each of which is capable of producing several gigabases of sequence data per run. As with the Roche 454, the ABI SOLiD exploits emulsion PCR to clonally amplify single template molecules onto beads. The beads are then attached to a glass surface, and millions of beads are sequenced in parallel by ligation (reviewed in Mardis, 2008).

In contrast, with the Illumina sequencing platform, molecules of DNA are hybridized to oligonucleotides that are attached to the polymer-coated glass surface of a flowcell (Fig. 18.2.1A). Templates are amplified by flowing enzymes and reagents through the channels of the flowcell (Fig. 18.2.1B). Once amplified, these molecules form clusters of amplicons, each of which is derived from a single template molecule. Clusters are then used as templates for sequencing-by-synthesis using fluorescent reversible-terminator deoxyribonucleotides (Bentley et al., 2008).

Next-Generation DNA sequencing technology is in the process of revolutionizing genetics, by enabling us to design genome-wide and ultra-deep sequencing projects that, because of their enormity, would not otherwise be possible. We are approaching a situation where whole-genome sequencing of complex organisms will be routine, which will allowing us to gain a deeper understanding of the full spectrum of genetic variation and to define its role in phenotypic variation and the pathogenesis of complex traits.

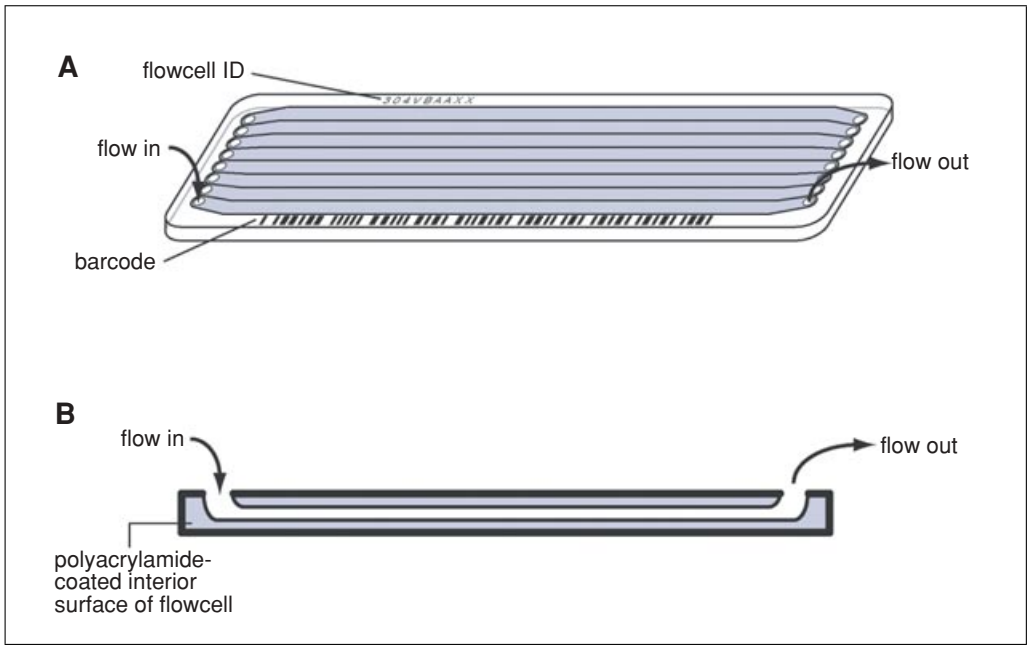

Figure 18.2.1 Illumina GAII flowcell. (**A**) Flowcells are hollow glass slides, with 8 separate lanes, through which reagents and template DNA flow. Lanes have been shaded gray for clarity. (**B**) Cross-section of a single lane, showing the direction of reagent flow and polyacrylamide coating on the interior surface of the flowcell.

QIAGEN Exhibit 1033.0002

In this unit, we describe in detail the molecular biology underpinning sequencing on the Illumina Genome Analyzer. We also describe a set of improvements we have made to the standard Illumina protocols to make the library preparation more reliable in a high-throughput environment, to reduce bias, tighten the distribution of fragment sizes, and to obtain high yields of data more reliably.

In order to achieve successful cluster amplification, sample DNA must undergo a multi-step library preparation procedure. We found several of the standard Illumina sequencing protocols could be improved upon, and by making modifications to their standard laboratory pipeline, we have made our sequencing output more robust and reproducible, and have also increased the output of our sequencing runs (Quail et al., 2008). The preparation steps are described below, and references to the protocols describing the optimized preparation steps are given.

Fragmentation. Cluster amplification is a relatively inefficient process, and there is an upper limit to the size of fragments that will amplify on the flowcell surface. Standard Illumina sequencing libraries currently tend to have a fragment size of 200 to 500 bp, excluding adapters (see Basic Protocol 1).

End-repair. Random fragmentation produces double-stranded DNA with a mixture of blunt ends, recessed 3′ ends and recessed 5′ ends, with and without a 5′ phosphate moiety. These must be made uniform before adapters can be ligated, and so a mixture of enzymes is used to generate blunt-ended fragments with phosphorylated 5′ termini.

A-tailing. Addition of a single A nucleotide to the 3′ ends of fragments before adapter ligation deters concatemerization of templates and increases the efficiency of adapter ligation.

Adapter ligation. Template strands must receive a different adapter sequence at either end to participate successfully in the cluster amplification and sequencing reactions. Adapter ligation to templates must be as efficient as possible, but at the same time, ligation of adapters to one another must be suppressed: adapter dimers will also generate clusters that can be sequenced, and will reduce the total proportion of desired sequence obtained from a run (see Basic Protocol 2).

Size selection. Ligated templates are run in an agarose gel, a gel slice is excised, and the DNA is extracted, to give a DNA library with a particular insert size range. Particularly for paired-end sequencing, it is advantageous to have as narrow and precise a range of fragment sizes as possible. Additionally, the gel step allows the majority of adapter dimers to be removed from the library (see Basic Protocol 2).

PCR. Following extraction from the gel slice, libraries are amplified by PCR (1) to enrich for properly ligated template strands—those that have an adapter at both ends, (2) to generate enough DNA for accurate quantification, and (3) to add oligonucleotide sequences to the template strands that allow hybridization to the flowcell surface, since these sequences are not contained in the ligated adapter (see Basic Protocol 3). For direct sequencing of short amplicons (without PCR), see Alternate Protocol 1.

Quantification. The number of clusters per lane of a flowcell is governed by the concentration of library that is added. Accurate quantification is vital, because too low a cluster density reduces the yield of data, and therefore increases the per-base cost of sequencing, whereas too high a cluster density results in a reduced yield of data due to cluster overlap (see Basic Protocol 4 or Alternate Protocol 2).

Denaturation. Libraries are rendered single stranded by incubation with sodium hydroxide, to allow efficient hybridization of the template strands to primers attached to the flowcell surface (see Basic Protocol 5).

QIAGEN Exhibit 1033.0003

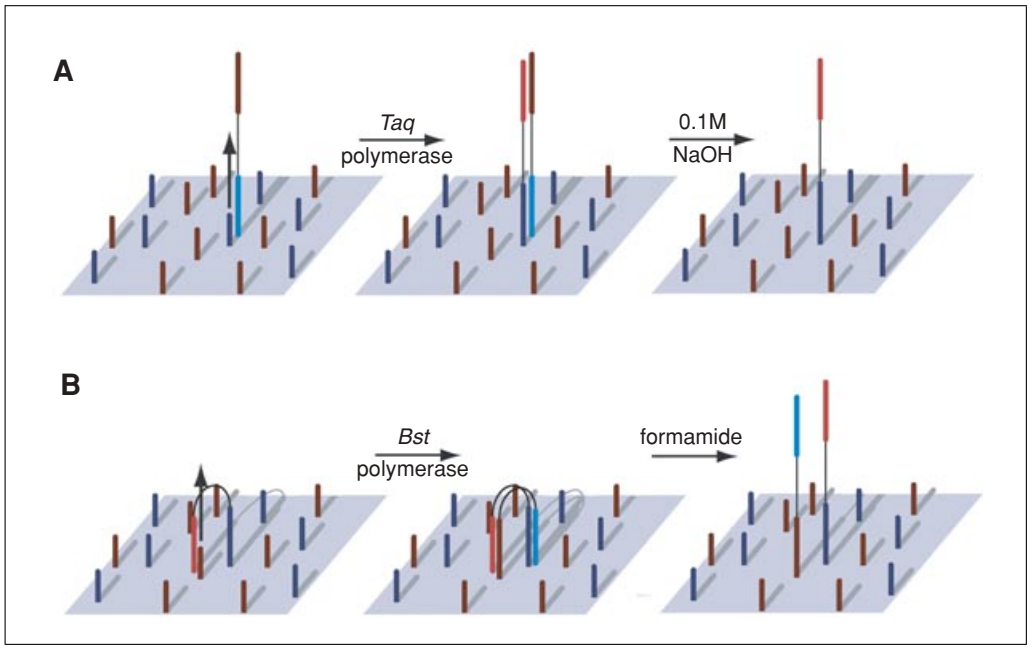

Figure 18.2.2 (**A**) Template hybridization, extension, and denaturation on the flowcell surface. Templates are prepared so as to possess tails that are complementary to primers on the flowcell surface. This allows one end of a template strand to hybridize to a flowcell primer. Flowcell primers are extended by *Taq* polymerase, resulting in a reverse complementary copy of the original template strand, which is covalently attached to the flowcell surface. The original template strand is then removed by flushing 0.1 M NaOH though the flowcell. (**B**) Cluster amplification. The free end of the tethered reverse complementary copy of the original template strand can anneal to the other type of flowcell primer, forming a bridge. The flowcell primer is extended by *Bst* polymerase, in an isothermal reaction, which generates a double-stranded product. Formamide is used to denature these strands, which can then anneal to other primers on the flowcell surface, which extend in the next cycle. In this way, repeated cycles of extension and denaturation result in a cluster of strands, all of which are derived from a single template strand.

Hybridization and extension. As a result of the PCR reaction, template strands possess a sequence at one end that matches exactly one of the flowcell primers, and at their opposite end they possess a sequence that is in the reverse complementary orientation to the other flowcell primer. Thus, only the complementary end of each template strand hybridizes to a flowcell primer. Flowcell primers are extended, by *Taq* polymerase, generating a reverse complementary copy of the original template strand that is tethered to the flowcell surface. The original template strand is not tethered, and is removed by flushing with sodium hydroxide (Fig. 18.2.2A).

Cluster amplification. The free end of each newly generated tethered strand is complementary to the other primer on the flowcell surface, and can hybridize to it. This primer is extended, using *Bst* polymerase, generating a double-stranded product. The strands of this product are denatured by formamide, producing two single strands, the free end of each of which can anneal to another primer on the flowcell surface (Fig.18.2.2B). Repeated cycles of formamide denaturation, annealing, and extension result in a cluster of ~1000 strands, each of which were derived from the same original template strand, and hence are clonal. (See Basic Protocol 6 for Amplification QC.)

Linearization, blocking, and hybridization. As a consequence of the structure of the adapters used for the ligation step, ligation junctions have around 12 identical nucleotides at opposite ends of the fragments. Therefore, to ensure that each cluster is sequenced only in one direction during each read, and to allow more efficient hybridization of sequencing primers to strands within a cluster, one of the flowcell primers is cleaved, removing one strand selectively, and resulting in single-stranded clusters. To reduce background noise

QIAGEN Exhibit 1033.0004

in the sequencing reaction, 3′ ends of both extended and unextended flowcell primers are blocked by incorporation of dideoxyribonucleotides, and sequencing primers are hybridized.

Sequencing-by-synthesis. Hybridized flowcells are transferred from the Cluster Station to the Genome Analyzer for cycles of nucleotide incorporation, imaging, and cleavage (simultaneous deblocking of the 3′ end and removal of the fluorophore).

FRAGMENTATION

BASIC PROTOCOL 1

The first stage in a standard genomic DNA library preparation for the Illumina Genome Analyzer is fragmentation of DNA by nebulization with compressed nitrogen or air. During nebulization, approximately half of the original DNA sample is lost through vaporization, so only 50% of the original DNA would be used for subsequent library generation. An additional drawback with nebulization is that it is not possible to shift the peak of fragment sizes much further towards the smaller end, even if more extreme conditions are used. With sonication, it is possible to produce fragments with a smaller peak. Nevertheless, standard sonication still produces a relatively wide range of fragment sizes, so a large proportion of the fragmented DNA is wasted.

We now routinely fragment all of our DNA samples using Covaris' Adaptive Focused Acoustics technology (AFA), which focuses acoustic energy controllably into the aqueous DNA sample. The peak of fragment sizes can be tuned to below 400 bp, so a greater proportion of the DNA sample will contribute to the final library. Moreover, a lower proportion of the sample is lost.

Materials

DNA sample
Qiagen QIAquick PCR purification kit (cat. no. 28104) containing:
- PB buffer
- PE buffer
- EB buffer
- Columns

Covaris S2 with chiller unit
Thermo Scientific NanoDrop
6-mm × 16-mm AFA fiber vials (100-µl; Covaris, cat. no. 520031)
Crimp caps (Covaris, cat no. 520028)

1. Allow the Covaris chiller to reach 4°C, and degas for at least 20 min.
2. During this time, prepare the DNA sample:
   a. Obtain an approximate concentration using a NanoDrop.
   b. Dilute 5 µg DNA to 100 µl with water and transfer the DNA sample to a 100-µl Covaris vial. Attach crimp cap by crimping.

      The aluminum lids perform better than the supplied plastic ones, as the tighter fit ensures that less acoustic energy is lost due to vibration.
3. Insert the sample vial into the holder, and for fragment sizes in the range of 200 bp, run the Covaris with the settings:

Duty cycle:	20%
Intensity:	4
Cycles per burst:	200
Time:	120 sec.

QIAGEN Exhibit 1033.0005

4. Clean up sample using a Qiagen QIAQuick PCR cleanup column, before proceeding with the end repair, eluting in 30 µl EB buffer. This concentrates the sample and removes the buffer in which the genomic DNA was originally suspended.

5. Proceed to template preparation (Basic Protocol 2).

BASIC PROTOCOL 2

TEMPLATE PREPARATION

Fragmentation generates templates with a mixture of blunt ends and 5′ and 3′ overhangs. So that adapters can be ligated, end repair and A-tailing reactions must be performed, following the manufacturer's protocols. For certain sequencing applications, such as surveying for rare translocation events, the frequency of chimeric templates must be kept to a minimum. In such cases, we perform an additional size selection, after the fragmentation step, and before end repair. In this way, any chimeric templates that do form in the ligation step will be far larger than the desired fragments and will be removed by the second gel step. For the majority of sequencing applications, a low frequency of chimeras is tolerable, and, in these cases, we perform a single-size selection step, after adapter ligation.

Materials

Fragmented DNA sample (see Basic Protocol 1)
Paired-end Sample Prep Kit (for library preparation; Illumina, cat. no. PE-102-1001) containing:
- 10× T4 DNA ligase buffer (+10 mM ATP)
- 10 mM dNTP mix
- T4 DNA polymerase
- Klenow DNA polymerase
- T4 polynucleotide kinase
- 10× Klenow buffer
- dATP
- 2× DNA ligase buffer
- DNA ligase

Qiagen QIAquick PCR purification kit (cat. no. 28104) containing:
- PB buffer
- PE buffer
- EB buffer
- Column

Klenow exonuclease (3′-5′ exo$^-$; Illumina)
Paired-end adapter mix (Illumina):
- PE_t_adapter: 5′ ACACTCTTTCCCTACACGACGCTCTTCCGATC*T (*indicates phosphorothioate; Bentley et al., 2008)
- PE_b_adapter: 5′ P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG (P- indicates phosphate; Bentley et al., 2008)

2% Invitrogen Ultra-pure agarose (Invitrogen, cat. no. 15510-027)
5× TBE buffer (Severn Biotech, cat. no. 20-6005-10)
10 mg/ml ethidium bromide solution (Sigma, cat. no. E1510)
Gel Pilot 5× loading dye (Qiagen, cat. no. 239901)
Low-molecular-weight size standard ladder (New England Biolabs, cat. no. N3233L)
Qiagen QIAquick gel extraction kit (cat. no. 28706) containing:
- Chaotropic buffer (QC buffer)
- Spin columns

37°C incubator
QIAquick MinElute column (included with the QIAquick MinElute kit; Qiagen, cat no. 28004)
Dark reader (Clare Chemical Research, cat. no. DR46)
Scalpel *or* razor blade

QIAGEN Exhibit 1033.0006

Perform end repair

1. Prepare the following reaction mix:

Fragmented DNA sample (Basic Protocol 1)	30 µl
Water	45 µl
10× T4 DNA ligase buffer (+ 10 mM ATP)	10 µl
10 mM dNTP mix	4 µl
T4 DNA polymerase	5 µl
Klenow DNA polymerase	1 µl
T4 polynucleotide kinase	5 µl
Total volume =	100 µl.

   Mix well and spin down.

2. Incubate 30 min at room temperature.

3. Purify with a Qiagen QIAquick PCR purification kit, following the manufacturer's instructions and elute in 32 µl of EB buffer.

Perform A-tailing

4. Prepare the following reaction mix:

End-repaired DNA sample (from step 3)	32 µl
10× Klenow buffer	5 µl
dATP	10 µl
Klenow exonuclease (3′-5′ exo$^-$)	3 µl
The total volume =	50 µl.

   Mix well and spin down.

5. Incubate 30 min at 37°C.

6. Purify on a QIAquick MinElute column, following the manufacturer's instructions, and elute in 10 µl of EB buffer.

Perform adapter ligation

7. Prepare the following reaction mix:

End-repaired, A-tailed DNA sample (from step 6)	10 µl
2× DNA ligase buffer	25 µl
Paired-end adapter mix	10 µl
DNA ligase	5 µl
Total volume =	50 µl.

   Mix well and spin down.

8. Incubate 15 min at room temperature.

9. Purify on a QIAquick column, following the manufacturer's instructions, and elute in 30 µl of EB buffer.

Perform size selection

10. Prepare a 2% Invitrogen Ultra-pure agarose gel in 1× TBE buffer, containing 0.4 µg/ml ethidium bromide (we typically use gels that are 12- to 15-cm long).

11. Following the ligation step, libraries are usually eluted in 30 µl EB buffer, so mix 30 µl DNA with 10 µl Gel Pilot 5× loading dye and run in the gel alongside a 100-bp-size standard ladder, in 1× TBE containing 0.4 µg/ml ethidium bromide at 6 V/cm, for ~2 hr (though this may need to be adjusted depending upon the length of the gel tank), until the yellow dye is close to the bottom of the gel.

QIAGEN Exhibit 1033.0007

If loading samples into a gel that is submerged in TBE buffer, run one sample per gel to remove the risk of cross-contamination.

Alternatively, it is possible to fill the gel tank so that the buffer level is just below the upper surface of the gel, and to load multiple samples. After sample loading, top up all wells with buffer (including empty wells), being careful not to overfill, and run the gel for 10 min, so that the samples enter the gel. Then remove the lid of the gel tank, add sufficient 1× TBE buffer (also containing 0.4 μg/ml ethidium bromide) to submerge the gel completely, and run as normal.

12. Visualize the gel on a dark reader.

    Ultraviolet light should not be used, as this can damage the DNA. It is helpful for visualization to make the room as dark as possible.

13. Using a clean scalpel or razor blade, cut a 2-mm gel slice, corresponding to the desired range of fragment sizes, allowing an additional 90 bp for the ligated adapters. Take care to cut horizontally so as not to increase the size ranges of fragments.

14. Using a Qiagen Gel Extraction kit, dissolve gel slices in chaotropic buffer at room temperature, rather than by heating. This typically takes 10 to 20 min with frequent mixing.

15. Continue purification using a QIAquick column, following the manufacturer's instructions, eluting in 30 μl EB buffer.

16. Proceed to PCR (Basic Protocol 3).

Perform double-size selection (optional)

17. After fragmentation, and column cleanup (Basic Protocol 1), run the entire column eluate in 4-wells of a 2% Invitrogen Ultra-pure agarose gel, alongside a 100-bp ladder.

18. For a 500-bp library, carefully take a horizontal slice corresponding to 400 to 600 bp.

    Following ligation, chimeric fragments will be 890 bp or longer (including adapters) and will be removed in the regular size selection step that follows. For smaller insert libraries, take a narrower gel slice, e.g., 200 to 225 bp.

19. Extract DNA from the gel slice as described above for the size selection (steps 14 and 15).

20. Proceed to PCR (Basic Protocol 3)

BASIC PROTOCOL 3

PCR AMPLIFICATION OF THE LIBRARY

Following size selection and gel extraction, libraries are amplified by 10 to 18 cycles of PCR to increase the quantity of the library, enrich for fully ligated fragments, and tail fragments with the nucleotide sequences necessary for cluster amplification. The number of PCR cycles is kept low, to limit amplification bias, and the mass of template DNA used is chosen for the specific library type (see below). This way, we can obtain a library that is clean and free from amplification artifacts. For this reason, we quantify ligated DNA using an Agilent Bioanalyzer DNA 1000 chip. When we are working with very dilute libraries, such as array eluates after a sequence capture experiment, we perform PCR using the protocol described below. For general purpose PCR amplification, we perform PCR using the manufacturer's reagents, following the recommended protocol.

QIAGEN Exhibit 1033.0008

Materials

Agilent DNA 1000 kit (cat. no. 5067-1504) containing:
- DNA dye concentrate (blue-capped vial)
- Gel matrix (red-capped vial)
- DNA marker (green-capped vial)
- DNA ladder (yellow-capped vial)
- Spin filters
- DNA chips
- Syringe

Extracted DNA library (see Basic Protocol 2)
Platinum *Pfx* polymerase and supplied 10× buffer (Invitrogen, cat. no 11708-013)
50 mM $MgSO_4$ (supplied with Platinum *Pfx* polymerase)
2.5 mM dNTP mix (Invitrogen, cat. no. R72501)
Paired-end PCR primers:
- PCR_F 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T (*indicates phosphorothioate; Bentley et al., 2008)
- PCR_R 5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T (*indicates phosphorothioate; Bentley et al., 2008)

Agencourt AMPure beads (cat. no. A29152)
70% ethanol

Vortex (supplied with the Bioanalyzer)
Agilent Bioanalyzer 2100
Thermal cycler
1.5-ml microcentrifuge tubes
Magnetic separator (e.g., Qiagen 12-Tube Magnet, cat. no. 36912)

Quantify DNA

1. To prepare the gel/dye mix, first allow dye concentrate (blue-capped vial) and gel matrix (red-capped vial) to reach room temperature (takes 30 min). After they reach room temperature vortex and spin down.

2. Pipet 25 µl of the dye concentrate into the gel matrix vial, cap the tube, vortex for 5 sec, spin down, and transfer the gel/dye to the supplied spin filter (supplied with the DNA 1000 kit). Centrifuge the spin filter 15 min at 2240 × g ± 20%, room temperature.

3. Store at 4°C, away from direct light when not in use for more than 2 hr.

4. Position a new chip onto the priming station, and pipet 9 µl of room-temperature gel/dye mix into the well marked G. Close the priming station, making sure that the syringe clip is in the lowermost position.

5. Press the plunger of the syringe until it is held by the clip and wait 60 sec.

6. Release the clip. If the plunger does not rise to ~0.6 ml within 5 sec, this suggests that the seal has failed. If so, replace the seal and repeat step 5.

7. Slowly pull the plunger back to the 1.0 ml position and open the priming station.

8. Pipet 9 µl gel/dye mix into both wells marked G.

9. Load 5 µl of marker (green-capped lid) into the well marked with the ladder symbol and each well that you are using to run a sample, and pipet 6 µl of marker into unused wells.

10. Load 1 µl of sample into each sample well and 1 µl of ladder into the ladder well.

QIAGEN Exhibit 1033.0009

11. Vortex on the supplied vortexer for 1 min and run the chip on the Agilent Bioanylzer 2100.

    To prevent excessive evaporation from the chip, the run should be started within 5 min after the vortexing. To quantify broad peaks, it is usually necessary to perform manual integration of the trace (right click on trace), to add a new peak and to extend the width of the new peak to include the area of interest.

Perform PCR

12a. For libraries that will be sequenced from a single end, use 3 ng DNA in a 50-µl PCR reaction, with 14 cycles of PCR amplification.

12b. For high-complexity paired-end libraries, such as those made from eukaryotic genomic DNA, use 25 ng DNA in a 50-µl PCR reaction with 12 cycles of PCR amplification.

12c. For lower-complexity libraries, such as bacterial genomic DNA, use 10 ng DNA in a 50-µl PCR reaction with 18 cycles of PCR amplification.

    These quantities give the optimal compromise between clean libraries and a low frequency of duplicate sequences.

13. In a 50-µl PCR reaction, mix 2 U Platinum *Pfx* polymerase in the supplied buffer diluted to 1× concentration, with 2 mM $MgSO_4$, 400 µM dNTPs and 1 µM primers, and the appropriate amount of DNA (see steps 12a to 12c).

14. Mix well and spin down.

15. Carry out the amplification using the following cycling conditions in a thermal cycler:

1 cycle	2 min	94°C	(initial denaturation)
10 to 18 cycles	15 sec	94°C	(denaturation)
	30 sec	62°C	(annealing)
	30 sec	72°C	(extension)
1 cycle	10 min	72°C	(final extension)
	indefinitely	4°C	(hold).

Clean up PCR

16. Mix the bottle of AMPure SPRI beads by gentle shaking. To a 50-µl PCR reaction, add 90 µl of bead suspension in a 1.5-ml microcentrifuge tube.

17. Mix by vortexing for 30 sec, to achieve a homogeneous suspension and leave at room temperature (20°C) for 5 min.

18. Spin down briefly and fix the tube into a magnetic separator. Leave until the solution is clear (~10 min).

19. Remove the cleared solution by pipetting, taking care not to disturb the beads.

20. Add 200 µl of 70% ethanol to the beads, taking care not to disturb the beads, leave for 30 sec at room temperature, remove the ethanol by pipetting, and discard the ethanol. Repeat this washing step once more.

21. Leave the lid of the tube open to allow the beads to dry.

    This can take longer than the recommended 20 min, but depends upon the volume of ethanol remaining after the previous step. The bead pellet will have a cracked appearance when dry.

22. Remove the tube from the magnet; add 40 µl water and mix thoroughly by pipetting. Be sure to resuspend the beads fully.

QIAGEN Exhibit 1033.0010

23. Return the tube to the magnet and leave for 10 min.

24. Collect and retain the liquid, as this now contains the DNA, and discard the beads.

25. Proceed to quantification (Basic Protocol 4).

DIRECT SEQUENCING OF SHORT AMPLICONS

ALTERNATE PROTOCOL 1

Unnecessary PCR amplification steps will exacerbate amplification biases, and so they should be avoided wherever possible. Extremely deep sequencing of short amplicons can be performed using locus-specific primers that possess tails capable of hybridization to the flowcell primers. The tail-less forward and reverse oligos can then be used as primers in the sequencing steps. Alternatively, sequencing primers can be chosen that anneal internally within the amplicon.

Materials

Genomic DNA template
2× hybridization buffer (see recipe)

Thermal cycler

1. Design tailed PCR primers:

   [AATGATACGGCGACCACCGAGATCTACACT] [specific primer 1]
   [CAAGCAGAAGACGGCATACGAGAT] [specific primer 2].

   This will enable the resulting amplicons to participate in the bridge amplification reaction.

   Instead of using the Illumina sequencing primers, use sequencing primers that match the locus-specific regions of the PCR primers. Locus-specific primer 1 will be the read 1 sequencing primer.

2. Quantify by qPCR, using Alternate Protocol 2.

3. Prepare a 2× hybridization buffer.

4. For read 1 (sequencing primer is hybridized to the flowcell on the Cluster Station) add 13.2 µl of 100 µM sequencing primer to 1306.8 µl 2× hybridization buffer. For read 2 (sequencing primer is hybridized to the flowcell on the Genome Analyzer) add 15 µl of 100 µM sequencing primer to 1500 µl 2× hybridization buffer and perform second-strand synthesis and read 2 primer hybridization following the manufacturer's instructions.

5. If using a control lane on the flowcell, it is necessary to use a mixture of sequencing primers during read 2, since these cannot be added separately to individual lanes of the flowcell. In this case, use 7.5 µl standard read 2 primer with 15 µl of 100 µM custom-sequencing primer in 1500 µl 2× hybridization buffer and perform second-strand synthesis and read 2 primer hybridization following the manufacturer's instructions.

6. Proceed to quantification (Basic Protocol 4).

QUANTIFICATION USING TaqMan PROBES

BASIC PROTOCOL 4

To obtain an optimal yield from a sequencing run, it is critical to quantify the library accurately before amplification on the flowcell surface. This is best achieved by quantitative PCR. The qPCR uses desalted forward and reverse primers that anneal to the flowcell-specific regions of library DNA, along with an HPLC-purified TaqMan probe, which anneals to the read 1 sequencing primer region of the library DNA. The

QIAGEN Exhibit 1033.0011

concentration of the template DNA is measured by comparison with a sequencing library that has been sequenced previously, and for which a precise cluster density is known.

Materials

TaqMan probe:
DLP (HPLC purified) [6FAM]CCCTACACGACGCTCTTCCGATCT[TAMRA]
Concentration standard (e.g., library previously sequenced with known cluster density)
10 mM Tris·Cl, pH 8.5 (*APPENDIX 2D*) + 0.1% Tween
50 mM $MgCl_2$ (supplied with Platinum *Taq*)
Template DNAs of unknown concentration (from Basic Protocol 3)
PCR primers:
c_qPCR_v2.1 (desalted) AATGATACGGCGACCACCGAGATC
PE_qPCR_v2.2 (desalted) CAAGCAGAAGACGGCATACGAGATC
50× Rox (Invitrogen, cat. no. 12223012)
2.5 mM dNTP mix (Invitrogen, cat. no. R72501)
Platinum *Taq* and supplied 10× buffer (Invitrogen, cat. no. 10966018)

Low-bind tubes (Axygen, MCT-175)
96-well qPCR plates (Applied Biosystems, cat. no. 4346906)
Adhesive plate sealers (Applied Biosystems, cat. no. 4311971)
Applied Biosystems StepOne Quantitative PCR machine (or equivalent)

1. Dilute all oligonucleotide solutions to a working concentration of 10 μM with water.

   The oligonucleotide solutions refer to the TaqMan probes. When these are purchased, they come as concentrated solutions, typically at 100 μM.

2. Chose a concentration standard that is as similar to the unknown sample as possible in terms of fragment-size range and base composition. This should be a library that you have sequenced previously, and for which you know the cluster density (or the total number of sequences per lane) from a given concentration (as measured by an Agilent Bioanalyzer 2100). Prepare three dilutions of this concentration standard: 100, 10, and 1 pM, based upon an Agilent Bioanalyzer 2100 quantification, diluting in 10 mM Tris·Cl, pH 8.5, + 0.1% Tween, and using low-bind tubes.

   NOTE: *It is essential to avoid any buffers containing EDTA, as this will inhibit the amplification reaction and will render quantification unreliable.*

3. Dilute template DNA of unknown concentration to 10 pM in 10 mM Tris·Cl, pH 8.5, + 0.1% Tween, also based on their Bioanalyzer concentration.

4. Perform qPCR assays in triplicate as follows (final concentrations are given in parentheses):

10× Platinum *Taq* buffer	2.5 μl
50 mM $MgCl_2$ (1.5 mM)	0.75 μl
Template DNA (from step 3) or concentration standard	2.5 μl
10 μM DLP (250 nM)	0.625 μl
50× Rox (1×)	0.5 μl
c_qPCR_v2.1 (300 nM)	0.75 μl
PE_qPCR_v2.2 (300 nM)	0.75 μl
2.5 mM dNTPs (200 μM)	2 μl
Platinum *Taq* (0.04 U/μl)	0.1 μl
H_2O	14.525 μl
Total volume:	25 μl.

QIAGEN Exhibit 1033.0012

5. Mix well and spin down. Cycling conditions are:

1 cycle:	2 min	94°C	(denaturation)
40 cycles:	15 sec	94°C	(denaturation)
	15 sec	62°C	(annealing)
	32 sec	72°C	(extension).

6. Adjust concentration of template DNA to 4 pM and proceed to denaturation (Basic Protocol 5).

QUANTIFICATION USING SYBRGreen PROBES

ALTERNATE PROTOCOL 2

Sequencing libraries that have been prepared using adapters that differ from the standard Illumina adapters (Alternate Protocol 1) cannot be quantified using the standard qPCR protocol (Basic Protocol 4), because both the TaqMan probe and the 3′ ends of the forward and reverse qPCR primers will be unable to anneal to the library. For such sequencing libraries, we use an alternate qPCR protocol, based on a SYBRGreen rather than a TaqMan probe, and which uses shorter primers. This protocol will also work with standard sequencing libraries.

Additional Materials *(also see Basic Protocol 4)*

PCR primers:
Syb_FP5 (desalted) ATGATACGGCGACCACCGAG
Syb_RP7 (desalted) CAAGCAGAAGACGGCATACGAG
Template DNAs of unknown concentration (from Alternate Protocol 1)
SYBRGreen PCR Master Mix (Applied Biosystems, cat. no. 4309155)

1. Dilute qPCR primers to a working concentration of 10 μM with water.

2. Prepare 100, 10, and 1 pM dilutions of a suitable concentration standard, as described above (Basic Protocol 4, step 2).

3. Prepare 1/100 and 1/10,000-fold dilutions of the template DNA that is to be quantified.

   Prepare a master mix of all reaction components apart from the template DNA, to ensure consistency. For example, if you have two templates of unknown concentration, each at 100- and 10,000-fold dilution, thus four samples, and three dilutions of concentration standard, you have a total of seven samples. Each will be assayed in triplicate, so you will have 21 qPCR reactions. Therefore, make a master mix for 25 reactions (an additional 20%) as follows:

2× SYBRGreen Master Mix	12.5 μl
10 μM Syb_FP5	0.75 μl
10 μM Syb_RP7	0.75 μl
Water	8.5 μl
Total volume:	22.5 μl.

   Mix thoroughly.

4. For each sample, prepare the triplicate qPCR reactions by adding 74.25 μl of master mix prepared in the preceding step to 8.25 μl of sample (gives three reactions plus an additional 10%), and mix thoroughly.

5. Pipet 25 μl of each reaction into a separate well of a 96-well PCR plate and seal with a plate seal.

QIAGEN Exhibit 1033.0013

6. Centrifuge the plate 1 min at 1200 × *g*, room temperature, and transfer to the qPCR machine. Cycling conditions are:

1 cycle:	10 min	95°C
40 cycles:	30 sec	95°C
	1 min	60°C.

7. Proceed to denaturation (Basic Protocol 5).

BASIC PROTOCOL 5

DENATURATION OF TEMPLATES

We denature all double-stranded DNA libraries in 0.1 M NaOH, before transferring an aliquot into 1 ml of hybridization buffer. It is essential not to transfer >8 µl of the denatured library into the hybridization buffer, because the pH becomes too high for efficient hybridization of the template DNA to the oligonucleotides on the flowcell surface.

Materials

2 N NaOH (Illumina)
Hybridization buffer (Illumina)
UltraPure water (Illumina)
Ice
DNA library (see Basic Protocol 3), concentration determined as in Basic Protocol 4
EB buffer (supplied with Qiagen QIAquick PCR purification kit, cat. no. 28104)

200-µl tubes
1.5-ml microcentrifuge tubes
Vortex

1. Make a 10-fold dilution of the supplied 2 N NaOH solution by adding 10 µl 2 N NaOH to 90 µl UltraPure water and mixing thoroughly.
2. Add 1 ml hybridization buffer into a 1.5-ml microcentrifuge tube and put on ice.
3. Dilute the DNA library to 2 nM using EB buffer.
4. Add 10 µl of the resulting 2 nM DNA library to 10 µl of the 0.2 N NaOH solution (prepared in step 1), rather than to 1 µl of the 2 N solution.

   This minimizes pipetting inconsistencies.

5. Vortex thoroughly and spin down.
6. Leave for 5 min at room temperature.
7. Transfer 4 µl to the hybridization buffer on ice (from step 2).
8. Proceed to cluster amplification.

   For cluster amplification, follow the manufacturer's protocol. After completing cluster amplification, proceed with the sequencing.

BASIC PROTOCOL 6

AMPLIFICATION QUALITY CONTROL

Following cluster amplification, DNA on the flowcell is double stranded and can be stained by an intercalating dye and detected on a fluorescence microscope. This is a useful quality control (QC) step, which we use for all flowcells prior to linearization and blocking to confirm that the cluster density is appropriate. We generally do not sequence flowcells that have too high or too low a cluster density (Fig. 18.2.3).

QIAGEN Exhibit 1033.0014

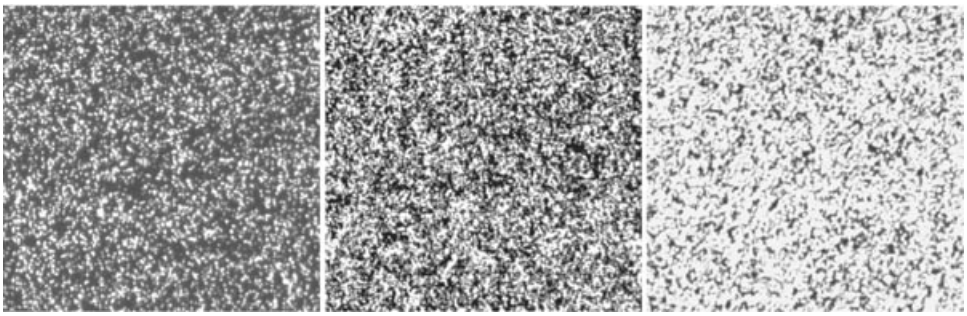

Figure 18.2.3 SYBRGreen QC. Although the most accurate method to measure cluster density is to perform a first-base incorporation on the flowcell, it is more economical to stain flowcells with SYBRGreen I immediately after amplification, and to examine cluster density qualitatively, using a fluorescence microscope. When coupled with qPCR quantification, this method is usually sufficiently accurate.

Materials

0.1 M Tris·Cl pH 8.0 (*APPENDIX 2D*)
Sodium ascorbate (Sigma, cat. no. A4034)
10,000× SYBRGreen I (Invitrogen, cat. no. 57567)
Amplified flowcell
PR2 buffer (Illumina, supplied with sequencing kits)

15-ml Falcon tubes
0.2-µm syringe filter
Cluster Station (Illumina)
Fluorescence microscope, set up to detect SYBRGreen I

1. Prepare 5 ml of a solution of 0.1 M Tris·Cl, pH 8.0, and 0.1 mM sodium ascorbate, and filter into a 15-ml Falcon tube using a 0.2-µm syringe filter.

2. Transfer 1960 µl of this Tris-ascorbate solution to a clean 15-ml Falcon tube, and add 40 µl of 100× SYBRGreen I (dilute from 10,000× stock solution with water), to produce a working concentration of 2×.

3. Using a hybridization manifold on the Cluster Station, manually flow 75 µl Tris-ascorbate per channel over 5 min through the flowcell.

4. Manually flow 150 µl Tris-ascorbate-SYBRGreen I per channel over 10 min.

5. Visualize clusters by fluorescence microscopy to check for an appropriate density.

6. Return the flowcell to the cluster station and flush through with PR2 buffer (150 µl per channel over 10 min), before storage or linearization and blocking.

   If the cluster density is out of the range of the 3 tiles shown in Figure 18.2.3, it may be uneconomical to proceed with the sequencing run.

QIAGEN Exhibit 1033.0015

REAGENTS AND SOLUTIONS

Use deionized, distilled water in all recipes and protocol steps. For common stock solutions, see ***APPENDIX 2D****; for suppliers, see* ***SUPPLIERS APPENDIX****.*

Hybridization buffer, 2×

Dilute 20× SSC + 0.2% Tween-20 stock solution (see recipe) with an equal volume of water (10× SSC + 0.1% Tween 20 final concentration) and mix thoroughly. Store up to 6 months at 4°C.

20× SSC + 0.2% Tween-20 stock solution

Dissolve the following in 800 ml distilled H_2O:
175.3 g NaCl (3 M final concentration)
88.2 g of sodium citrate (0.3 M final concentration)
2 ml Tween-20 [0.2% (v/v) final concentration]
Adjust the pH to 7.0 with 0.1 M HCl
Adjust volume up to 1 liter with distilled H_2O
Filter using a 0.2-μm vacuum filter
Store up to 6 months at 4°C

COMMENTARY

Background Information

The sequencing reaction on the Illumina Genome Analyzer platform takes place on the interior surfaces of a hollow glass slide, termed a flowcell, which is approximately the same size as a standard microscope slide. A flowcell is divided physically into eight lanes (Fig. 18.2.1A), allowing up to eight different sequencing libraries to be sequenced in a single run. Sequencing libraries consist of a collection of DNA fragments, with a specific range of sizes, which are ready to be sequenced. The interior surfaces of a flowcell are coated in polyacrylamide (Fig. 18.2.1B), to which two oligonucleotides are attached, creating a random lawn of both oligos (Fig. 18.2.1, panels A and B). These act as forward and reverse primers for the exponential, isothermal cluster amplification reaction, which is performed by repeated cycles of extension, denaturation, and annealing on a Cluster Station. Because primers are attached to the polyacrylamide covalently, cluster amplicons are tethered to a fixed position on the flowcell surface. Amplified clusters consist of double-stranded DNA, and one strand is removed selectively before sequencing.

The flowcell is then transferred to a Genome Analyzer, where clusters undergo a sequencing-by-synthesis reaction using reversible fluorescent terminator deoxyribonucleotides. Being terminator nucleotides, each DNA strand within a cluster can only incorporate a single nucleotide during each chemistry cycle, and being clonal, each strand within a cluster incorporates the same nucleotide. Clusters are imaged, blocking groups and fluorophores are removed by chemical cleavage, and the next round of nucleotide incorporation begins. Images are analyzed, generating a separate sequence for each cluster. Sequence length is identical for all clusters, as it is governed by the number of cycles of nucleotide incorporation, imaging, and cleavage.

Library preparation

The purpose of the library preparation reactions is to introduce adapter sequences onto template molecules that allow amplification onto the flowcell surface. Here we have described a number of modifications that allow for more efficient sample preparation, and which enable a stable workflow in a production environment.

Fragmentation

The first stage in a standard genomic DNA library preparation for the Illumina Genome Analyzer is fragmentation of DNA by nebulization with compressed nitrogen or air. This is performed over 6 min, in 30% to 60% glycerol at 30 to 35 psi and generates fragments with a typical size range of 0 to 1200 bp and a peak around 5 to 600 bp. Nebulization is a fairly reproducible technique, and is sequence-independent, rapid, and inexpensive (Surzycki, 2000). However, the range of fragment sizes generated by nebulization is wide. Sequencing libraries are typically prepared with a narrow range of insert sizes, so the majority of fragments will be wasted, which increases the amount of sample DNA needed at the beginning of the process. For example,

QIAGEN Exhibit 1033.0016

a typical small insert-sequencing library has a fragment size range of 180 to 220 bp, which constitutes ~10% of the total DNA by mass after nebulization. During nebulization, approximately half of the original DNA sample is lost through vaporization, so in this example, only 50% of the original DNA would be used for subsequent library generation.

An additional drawback with nebulization is that it is not possible to shift the peak of fragment sizes much further towards the smaller end, even if more extreme conditions are used. For example, a gas pressure of 60 psi for 12 min produces a peak at ~450 to 500 bp, but further increases in pressure or time only reduce yield (Surzycki, 2000; and personal observation). As a consequence, we have evaluated alternative methods of sample fragmentation.

Sonication has the advantage over nebulization that the peak of fragment sizes can be tuned to below 400 bp, so a greater proportion of the DNA sample will contribute to the final library. Moreover, a lower proportion of the sample is lost. However, like nebulization, sonication still produces a relatively wide range of fragment sizes, so a large proportion of the fragmented DNA is wasted.

We now routinely fragment all of our DNA samples using Covaris' Adaptive Focused Acoustics technology (AFA). Here, acoustic energy is focused controllably into the aqueous DNA sample by a dish-shaped transducer, which creates cavitation events within the sample. The collapse of bubbles in the suspension creates multiple, intense, localized jets of water, which disrupt the DNA molecules in a reproducible and predictable way.

Following disruption, 200-bp fragments comprise 17% of the total fractionated DNA by mass, but in contrast to nebulization, very little DNA is lost during the fragmentation process, generating a 4- to 5-fold higher yield of the intended fragment size range than nebulization (Fig. 18.2.4). In addition, because the size distribution of DNA fragmented by AFA is narrow, particularly with the newer 100-µl (6 mm × 16 mm) vials that contain AFA fiber, for some applications, such as array enrichment of targeted loci (Albert et al., 2007; Hodges et al., 2007), we can omit the gel size selection step altogether from the library preparation, decreasing the workload and increasing yields further.

A-tailing, adapter ligation, size selection, and gel extraction

Analysis of paired-end sequence data from the Genome Analyzer—in which each cluster was sequenced in both forward and reverse directions—revealed several artifacts that could be attributed to the standard library preparation protocol:

1. *Bias in the base composition of sequences:* The mean GC content of the sequences obtained differed from that of the organism from which the sequences were derived.

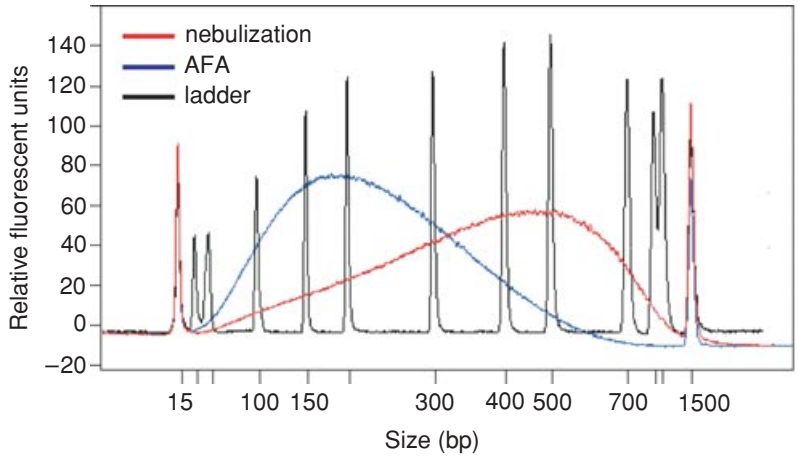

Figure 18.2.4 Comparison of sample fragmentation by nebulization with Covaris AFA. 4.5 µg human genomic DNA was fragmented by nebulization (red line) and AFA (blue line). Both were purified using a spin column and eluted in 30 µl EB buffer (Qiagen). 1 µl of each eluate was run on an Agilent Bioanalyzer DNA 2100 chip. Image adapted with permission from Macmillan Publishers Ltd. (Quail et al., 2008). For color version of this figure go to *http://www.currentprotocols.com/protocol/hg1802*.

QIAGEN Exhibit 1033.0017

2. *High frequency of chimeric sequences:* These are sequences for which the two paired-end reads map to regions of the genome that are separated by far more than the intended insert size. Though this could indicate a genuine deletion or translocation in the sample, and could be confirmed by PCR, a high frequency of chimeric sequences is most likely to be a library preparation artifact.

3. *Imperfect distribution of insert sizes:* A perfect distribution should be Poisson-like, with a peak at the expected position.

These artifacts have been overcome by the use of several protocol modifications.

A-tailing and adapter ligation

Paired-end libraries can be amplified and sequenced on both paired- and single-end flowcells, though of course, a paired-end library on a single-end flowcell can still only be sequenced in one direction. We generally prepare all of our libraries to be paired end, as this gives us more flexibility: a library that was first run on a single-end flowcell can be rerun on a paired-end flowcell without repeating the library prep.

Prior to adapter ligation, templates are given an A-overhang on the 3′ end of each strand, which complements a 3′ T-overhang on the adapter. This makes ligation more efficient than if it were blunt ended. A-tailing also hinders blunt-ended self ligation of templates, which would otherwise generate chimeric sequences. The ligation adapters themselves are modified on one strand using a phosphorothioate modification between the T-overhang and the penultimate base at the 3′ end (Bentley et al., 2008). This prevents removal of the T-overhang by any contaminating exonuclease activity in the ligase preparation, which prevents blunt-ended self-ligation of adapters. The other strand is phosphorylated at the 5′ end, allowing efficient ligation to templates.

Both single- and paired-end adapters are partially complementary, so the end that ligates to the template is double stranded, whereas the opposite end is not. Essentially, the adapters consist of the nucleotide sequences to which the sequencing primers hybridize during the sequencing-by-synthesis reaction. These are ligated onto the A-tailed fragments (Sambrook et al., 1989) via their T-overhang. Their structure ensures that each template strand receives different sequences at the opposite ends (Smith and Malek, 2007; Bentley et al., 2008), and works in a similar way to a vectorette (Riley et al., 1990).

Size selection

Despite all of the preventative measures used above, adapter dimers do still form during the ligation step. This is possibly due to some remaining exonuclease activity, although the sequences obtained appear to be more consistent with the T-overhangs annealing to one another. Adapter dimers should be removed, so as not to waste the sequencing capacity of a flowcell. Running ligated samples in an agarose gel and excising a band is a convenient way of achieving this, and at the same time, allowing fragments of a defined insert size to be selected.

Gel extraction

Although excision of a gel slice is standard only in the single-end library prep protocol, we found that taking a 2-mm gel slice for PE libraries, rather than performing a gel stab, greatly improves the robustness of the library prep. We identified that melting this gel slice by heating to 50°C in Qiagen's QG buffer decreased the representation of A/T-rich sequences, possibly reflecting a higher affinity of spin columns for double-stranded DNA, as strands with a high A/T content will be most likely to become denatured during this step, and least likely to re-anneal. To improve the representation of these A/T-rich sequences we modified the gel extraction protocol, melting agarose gel slices in the supplied buffer at room temperature, and found this to reduce GC bias considerably (Fig. 18.2.5A,B).

Double size selection

Template molecules that have not been A-tailed at the 3′ ends of both strands possess one or two blunt ends, and so are substrates for blunt-ended ligation. This results in a chimeric template molecule. Because ligation is performed before any size selection step, the full range of fragment sizes will be present. If, for example, two blunt 100-bp fragments ligate together, and if the desired fragment size is 200 bp, during the gel size selection step the chimera will be excised and extracted along with the fragments that are genuinely that size. For many sequencing applications, a low frequency of chimeric sequences can be tolerated, and can be removed informatically, as they will map to distant parts of the genome. For other applications, such as screening for translocations, these in vitro translocations can be falsely interpreted as genuine structural variants, and they will require a larger amount of subsequent confirmatory work.

QIAGEN Exhibit 1033.0018

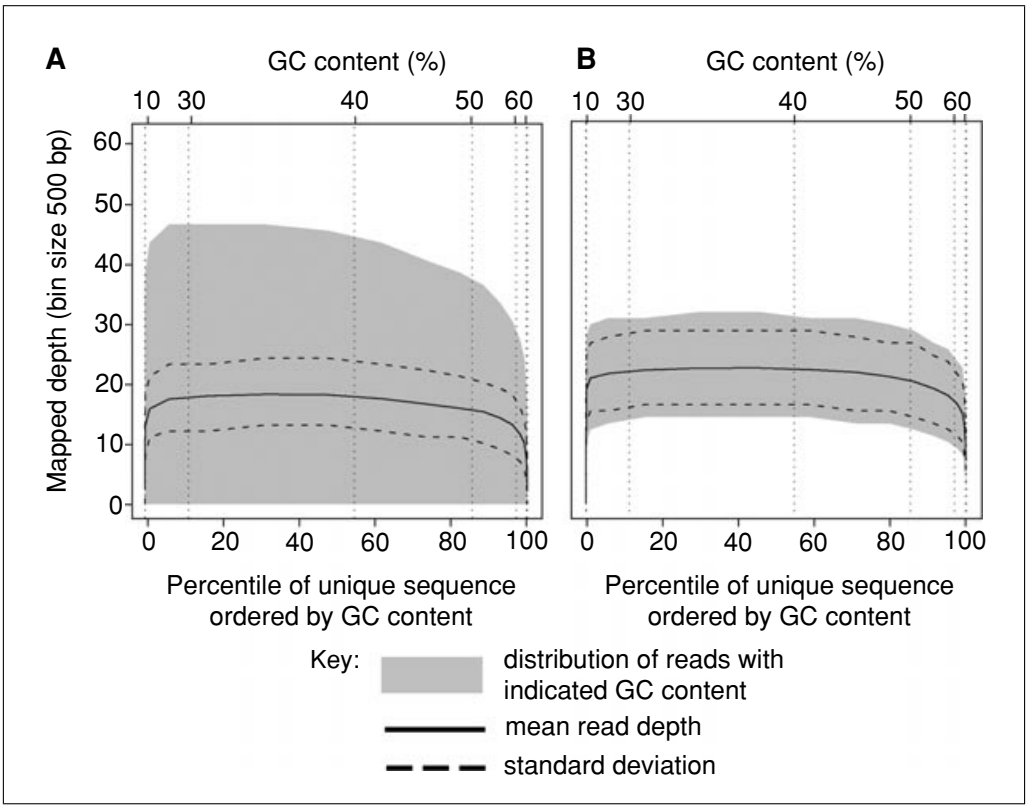

Figure 18.2.5 Comparison of gel extraction with and without heating. The plots show the total area in which reads with a particular G+C content are distributed; the mean and standard deviation are also shown. (**A**) This plot represents the standard gel extraction protocol, in which gel slices are heated to 50°C. (**B**) This plot shows the G+C distribution for the optimized gel extraction. The greater width of the shaded area in plot (A) indicates a wider dispersion of coverage for all values of G+C content for which sequences were obtained. Image adapted with permission from Macmillan Publishers Ltd. (Quail et al., 2008).

The frequency of such chimeric templates can be reduced by performing an additional size selection immediately after fragmentation and purification of the sample DNA. Consequently, most of the chimeric templates will fall out of the size range of the second size selection. This additional size selection reduces the incidence of chimeras from ~5% to 0.02%, and we have found the step to have the added benefit of reducing the shoulder of fragments with small insert sizes, which is sometimes evident (Fig. 18.2.6A,B), giving a tighter insert size distribution of the desired fraction, which leads to clusters with more uniform diameter.

PCR

To produce a clean sequencing library, it is advantageous to use optimized quantities of template in the PCR. We routinely analyze our post-PCR sequencing libraries by performing microchip capillary electrophoresis, using an Agilent Bioanalyzer 2100. This allows us to quantify the library, but also to detect products that differ in size from the expected amplicons. In this way, we noticed that the quality of a post-PCR library decreases as the amount of template DNA used in the PCR increases: too much template DNA often results in the generation of an apparently higher molecular weight peak (Fig. 18.2.6C). This is typically twice the size of the expected product, as measured by the Bioanalyzer, and may represent a single-stranded template product that accumulates as primers become depleted. Conversely, if too little DNA is used in the PCR, the smaller the pool of original templates, and the greater the incidence of PCR duplicates in the resulting sequences. PCR duplicates are pairs of sequences for which sequences map to identical positions in the genome. Some duplicate sequences will inevitably arise by chance, when two molecules in the sample are sheared at the same position at both ends, but the frequency of this is very low and predictable, and depends upon the read length and depth of sequence coverage. A low frequency of duplicate sequences (~0.1%) also arises by the cluster detection software misinterpreting single clusters as pairs. However, the vast majority of observed duplicates arise during the PCR.

QIAGEN Exhibit 1033.0019

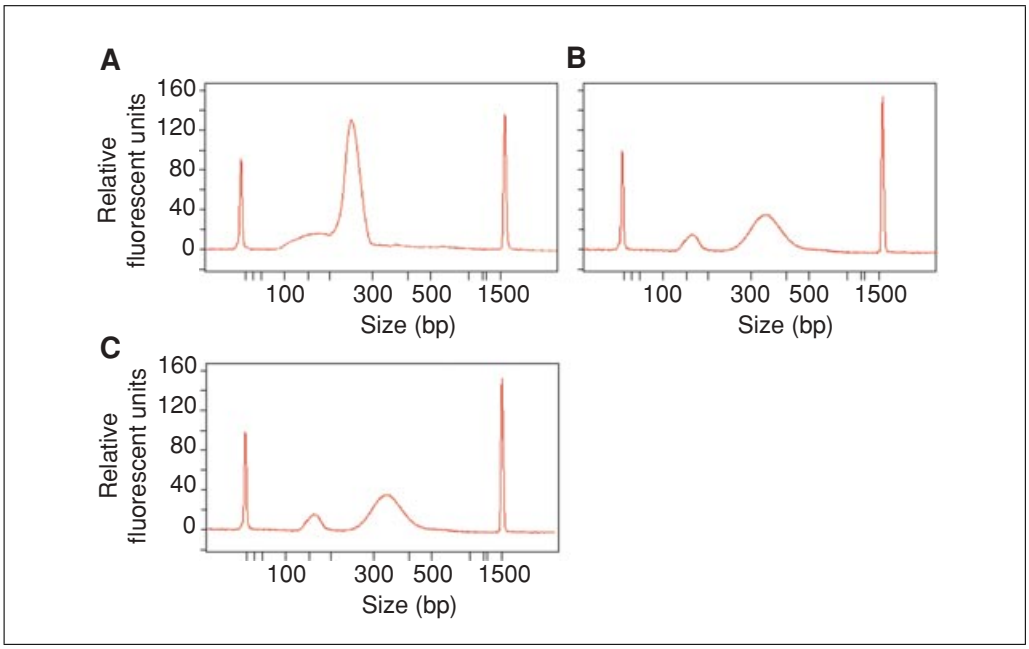

Figure 18.2.6 Size selection and PCR. Agilent Bioanalyzer DNA 1000 traces for three libraries (**A**) a double-size selected 100-bp insert library that was amplified using optimized PCR conditions, (**B**) a 200-bp insert library (single size selection) showing a shoulder of smaller fragments, (**C**) the same double-size selected 100-bp insert library as (A) but using standard PCR conditions. The peaks at 15 and 1500 bases are Agilent-supplied size standards. Image adapted with permission from Macmillan Publishers Ltd. (Quail et al., 2008).

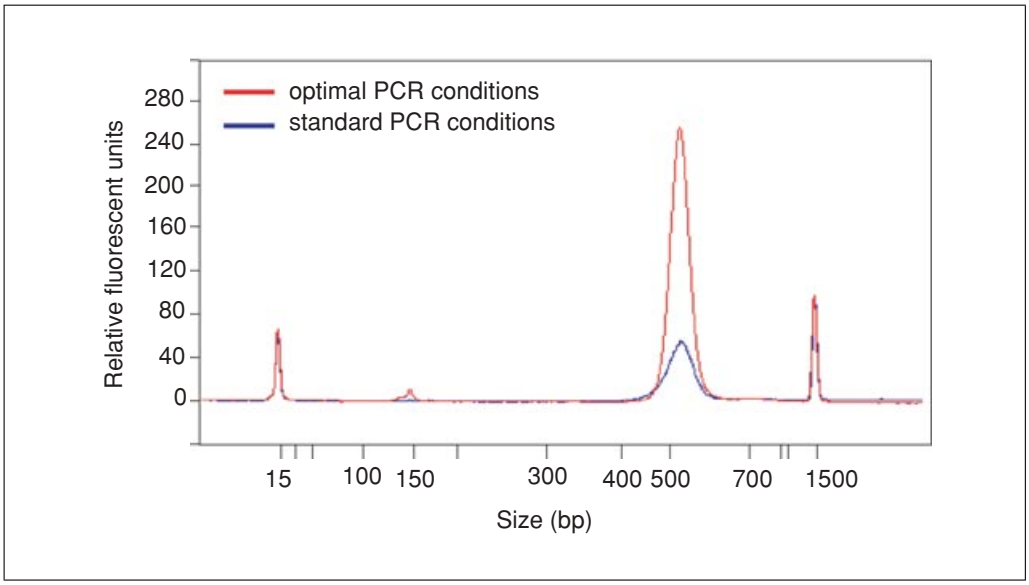

Figure 18.2.7 Increased PCR yield using improved conditions. A 500-bp library was prepared, and 1 ng was amplified for 18 cycles of PCR using standard conditions (blue curve) and our optimized conditions (red curve). Image adapted with permission from Macmillan Publishers Ltd. (Quail et al., 2008). For color version of this figure go to *http://www.currentprotocols.com/protocol/hg1802*.

Thus, it is essential to choose the appropriate set of conditions for each PCR.

PCR yield

The standard Illumina PCR uses Phusion polymerase and a premixed buffer, but by using alternative high-fidelity polymerases and optimizing the reaction further, we have been able to increase the yield of the enrichment PCR reaction 5- to 10-fold (Fig. 18.2.7), which allows fewer cycles of amplification to be performed.

PCR cleanup

Surplus PCR primers may interfere with quantification and will compete with the

QIAGEN Exhibit 1033.0020

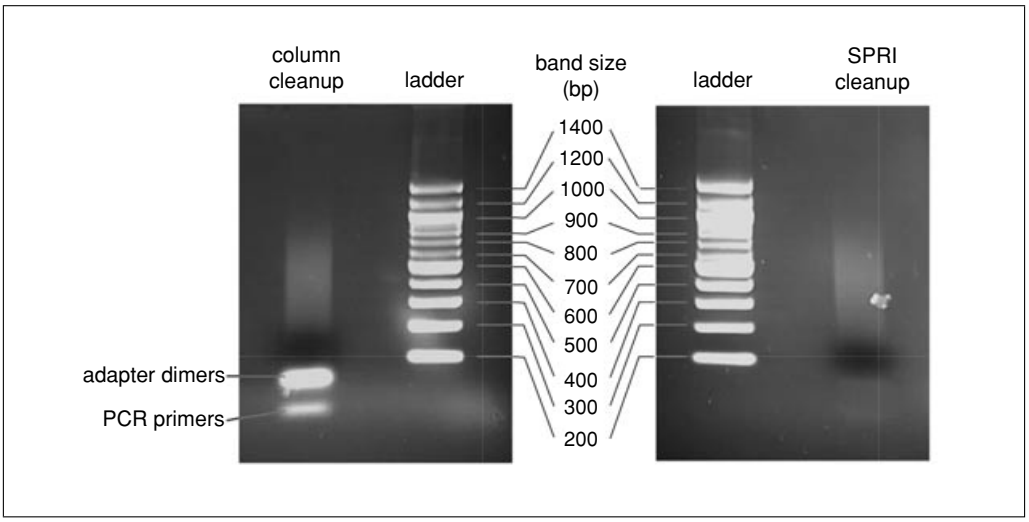

Figure 18.2.8 PCR cleanup. We prepared a paired-end PhiX library using conditions that would promote the formation of adapter and primer dimers and unextended PCR primers. After PCR, we divided the library in two: half was purified using a QIAquick spin column, as in the standard Illumina protocol (left), whereas the other half was purified using AMPure SPRI beads (right). Gels are shown after staining and excision of the gel slice corresponding to the desired size range of fragments. Image adapted with permission from Macmillan Publishers Ltd. (Quail et al., 2008).

amplicon for hybridization to the flowcell surface, but a more significant problem is presented by adapter dimers that enter the PCR reaction. Here, dimers also receive the full-length nucleotide tails that allow hybridization and amplification on the flowcell surface, and so will form clusters that are sequenced alongside the desired templates. Consequently, it is advantageous to remove dimers, as well as unextended oligos after the PCR. We have found that solid-phase reversible immobilization (SPRI) technology (Hawkins et al., 1994) removes a higher proportion of primers and adapter dimers than spin columns, without compromising on yield, and also allows elution in a wider variety of buffers (Fig. 18.2.8).

Library quantification

Accurate quantification of DNA prior to cluster amplification is essential. For fragment sizes undergoing a given number of cycles of cluster amplification, there is a concentration range of DNA that will yield clusters in the optimal density range, enabling the maximum amount of data to be obtained (Fig. 18.2.9). For fragments with a mean insert size of 500 bp or lower we aim for ~180,000 clusters per imaged area (=tile) on the GAII, giving ~140,000 purity filtered (PF) clusters per tile, equating to 4.0 Gb per 37-cycle single-end run. It should be noted that optimal cluster densities are dependent upon which version of the Illumina pipeline analysis software is run.

Electrophoresis (Agilent Bioanalyzer 2100)

Cluster density based on spectrophotometry tends to be inconsistent, but typically 5- to 10-fold lower than expected for a given library concentration, presumably because spectrophotometry cannot distinguish between differently sized DNA species, and measures not only the intended amplicon but also adapter dimers and unextended primers. Spectrophotometry also struggles to measure low DNA concentrations accurately.

Using an Agilent Bioanalyzer 2100 for library quantification, we can achieve a much more consistent cluster density. Additionally, because the Agilent can determine the size of DNA species, it allows us to check the quality of the sample preparation. In spite of this, however, for a small proportion of libraries, we obtained far higher cluster densities, and consequently far less useful data, than the measured concentration value would predict. We assume that this is a result of single-stranded DNA generated in the PCR: the Agilent Bioanalyzer cannot quantify single- and double-stranded DNA together. Although optimized PCR conditions can help us to avoid the generation of single-stranded DNA, we also sought to develop a quantification assay that could detect all amplifiable template molecules in a library.

Quantitative PCR

Quantitative PCR should be capable of detecting and quantifying all amplifiable

QIAGEN Exhibit 1033.0021

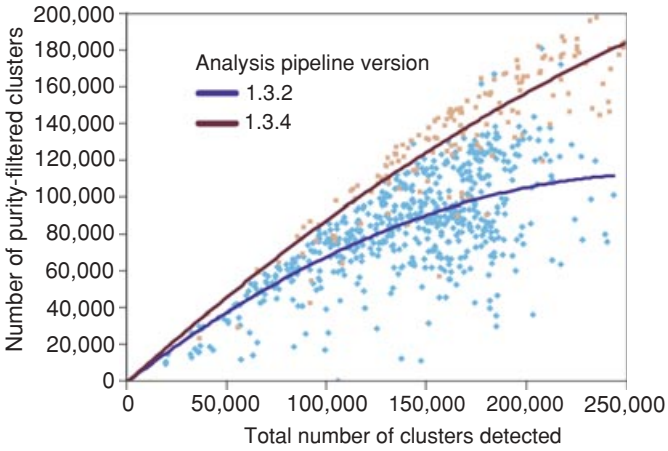

Figure 18.2.9 Cluster throughput as a function of total clusters. The graph shows analyzable purity filtered (PF) clusters in one tile (imaged area) versus total raw clusters per tile. The data was obtained from GAII flowcells using two different versions of the analysis pipeline, 1.3.2 and 1.3.4. For version 1.3.2, it can be seen that above the optimal cluster density, the number of clusters obtained after purify filtering begins to decrease as more and more clusters overlap, and so are discarded by the image analysis software. With pipeline version 1.3.4, the relationship between total and PF clusters is more linear. In both cases, accurate library quantification is essential for the sequencing run to generate the maximum yield of data.

molecules—i.e., those with adapters at either end (for discussion see Meyer et al., 2008). We designed amplification primers and a dual-labeled (TaqMan) probe to anneal to the Illumina paired-end adapter sequences. Because the amplification of Illumina libraries is rarely 100% efficient, we quantify unknown libraries against a dilution series of a concentration standard. This is a library that has been sequenced previously, and for which we know the accurate cluster number, and how this relates to the concentration of that library as measured by the Agilent Bioanalyzer. A concentration standard is also typically a library that has a similar base composition and insert size range to the unknown library. This allows us to predict cluster number accurately (Fig. 18.2.10).

Denaturation

After quantification, double-stranded DNA libraries are denatured with 0.1 M NaOH before diluting and loading onto the flowcell. Because a high pH prevents efficient hybridization, Illumina recommends that no more than 8 µl of denatured library be added to 1 ml hybridization buffer, to avoid carryover of excess NaOH. Given that the optimal loading concentration is ~4 pM, denaturation of libraries that are below 0.5 nM is problematic, as these require >8 µl of library to be added to the hybridization buffer: denaturation by heating has the potential both to damage the DNA and to introduce anti-GC bias (Mandel and Marmur, 1968). Consequently, we still prefer to denature dilute templates with NaOH, using Basic Protocol 5.

Critical Parameters

Fragmentation and post-fragmentation QC

Prior to working on real samples it is recommended that a series of fragmentation experiments be performed upon various concentrations of a test DNA sample (e.g., human genomic DNA; Promega, cat. no. G1471). After fragmentation, results can be visualized by running the sheared DNA on an agarose gel or Agilent Bioanalyzer DNA 1000 chip allowing the identification of optimal shearing parameters that give a maximal proportion of DNA within the desired size range.

The starting quantity of DNA can have a critical effect on the success of library preparation. For standard paired-end libraries, we use at least 500 ng, though we recommend that 5 µg genomic DNA be used if performing double size selection. Quantification of genomic DNA tends to be unreliable, and can lead to suboptimal amounts of DNA being available for library preparation. We do not encourage the use of spectrophotometric methods for quantification as these can be rendered inaccurate by small nucleic acids and contaminating chemicals; we have found SYBRGreen-based assays to be more trustworthy (e.g.,

QIAGEN Exhibit 1033.0022

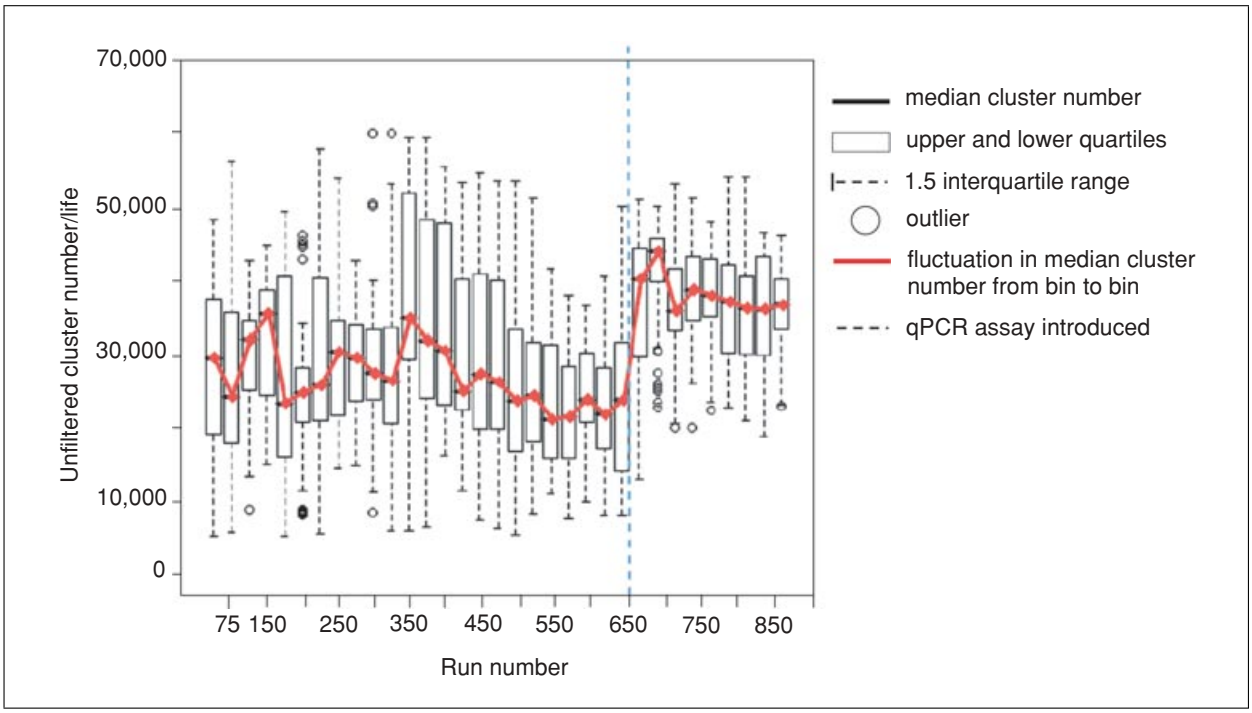

Figure 18.2.10 Improvement of cluster density reproducibility with qPCR quantification. Runs were grouped into 25-run bins, and a boxplot was generated. After some initial problems with degradation of standards, cluster number leveled out at ~35,000 to 40,000 per tile for GAI flowcells. Image adapted with permission from Macmillan Publishers Ltd. (Quail et al., 2008). For color version of this figure go to *http://www.currentprotocols.com/protocol/hg1802*.

Invitrogen Qubit fluorimeter—order code Q32857), though concentrations can be overestimated in samples containing excessive amounts of contaminating RNA.

If sufficient genomic DNA (>2 μg) is available, the sample can be analyzed after fragmentation on an Agilent Bioanalyzer DNA 1000 chip to assess the size distribution that has been generated, and to confirm quantification. If necessary, samples that have not sheared successfully can be subjected to further fragmentation.

End-repair, A-tailing, and adapter ligation

These reactions are generally very robust, although it is essential to ensure that buffers have been thawed completely, and that any particulate matter has fully dissolved. To achieve this, leave buffers, oligos, and adapters at room temperature (20°C) for 30 min prior to use. Check for any precipitated material by visual inspection, and if this is present, warm the buffer to 37°C, vortex, and spin down. Avoid repeated freezing and thawing of these buffers. Store enzymes at −20°C and take out of the freezer, spin down, and place tubes on ice just before use. Return enzymes to −20°C immediately after use.

When setting up reactions, we recommend using a tick list to record when each component has been added. Enzymes should be added last, and all reactions should be mixed well by gently pipetting up and down three times.

Pre-PCR QC

The quantity of template used in the PCR can affect library quality. We quantify the adapter-ligated template prior to PCR using a Qubit HS assay (Invitrogen).

If template concentration is too low to be measured, either insufficient genomic DNA was used to begin with, or too much loss has occurred during the column cleanup steps. In either event, it is necessary to repeat the library prep. To enhance yields from column cleanup steps, ensure that the PBI buffer is mixed thoroughly with the sample before adding to the column, and that the ethanol has evaporated after the wash step (leaving at 37°C for 15 min is sufficient), and wait for 5 min after adding EB buffer to the column before eluting.

It is also possible to quantify the amount of adapter ligated DNA at this stage by qPCR to determine what fraction of the library has adapters on both ends.

QIAGEN Exhibit 1033.0023

Table 18.2.1 Troubleshooting Guide for Preparing DNA Libraries Used in the Illumina Sequencing System

Problem	Possible cause	Solution
Insufficient DNA after fragmentation	Inaccurate quantification of genomic DNA	Fragment additional genomic DNA
	Poor recovery from cleanup column	Mix sample and buffer thoroughly before adding to column; allow column to dry thoroughly after rinsing with wash buffer; incubate column with elution buffer for 5 min before eluting.
Unexpected size distribution after fragmentation	Inappropriate settings used on Covaris. Incorrect vials used on Covaris.	Check settings and vials
Insufficient DNA after adapter ligation (<25 ng total)	Poor recovery from cleanup column/gel	Mix sample and buffer thoroughly before adding to column; allow column to dry thoroughly after rinsing with wash buffer; incubate column with elution buffer for 5 min before eluting.
Large peak at 130 bp after PCR	Adapter dimer contamination—size range of selected fragments not optimal	Cut gel slice corresponding to larger fragment size, or purify with AMPure beads using a DNA:bead ratio of 1:0.9.
Insufficient DNA after PCR (<2 ng/µl)	PCR failure	Repeat PCR with fresh reagents
	Fragment size range too broad	Perform qPCR quantification on stock and 1/10 dilution
No clusters visible in lane 8, column 2	Insufficient oil added to right-hand side of flowcell	Add more oil to right-hand side of flowcell
Clusters blurred	Poor focusing	Repeat focusing steps
	Oil on flowcell	Clean flowcell gently with lens tissue and ethanol
Low cluster intensity	Failure of linearization or primer hybridization	Repeat linearization/block/hybridization
Low % of PF clusters	Cluster density too high (>200,000/tile)	Repeat amplification with lower concentration
	Too wide a range of fragment sizes	Repeat size selection, taking narrower gel slice
Low number of raw clusters detected	Cluster density too high (>200,000/tile)	Repeat amplification with lower library concentration
	Cluster density too low (<100,000/tile)	Repeat amplification with higher library concentration
	Degradation of concentration standard used in qPCR	Repeat qPCR with fresh dilutions of concentration standard
Spiky IVC plots	Adapter contamination	If % adapter is too high (e.g., >10%) repeat size selection step of library prep

continued

QIAGEN Exhibit 1033.0024

Table 18.2.1 Troubleshooting Guide for Preparing DNA Libraries Used in the Illumina Sequencing System, *continued*

Problem	Possible cause	Solution
High A signal in basecall plots	Low cluster density	Repeat amplification with higher library concentration
Anti-AT bias in sequences	Heating during gel step	Dissolve gel slice at room temperature
	Too many PCR cycles	Repeat using a maximum of 10 cycles
High % duplicate sequences	Too little ligated DNA in PCR	Repeat PCR with higher quantity of ligated DNA
	Inefficient end repair/ A-tailing/adapter ligation	Repeat library prep using fresh reagents

Post-PCR QC

A 1-µl aliquot of each library should be run on an Agilent Bioanalyzer DNA 1000 chip to check concentration and gauge library quality. This reveals the amount of adapter dimer present in the library (observed as a sharp peak at ∼130 bp), and the distribution of fragment sizes present in the library.

Quantification

Although the Agilent Bioanalyzer is useful in determining approximate library concentrations, it cannot be relied upon completely, and the concentration measurement should be considered an initial estimate, which allows samples to be diluted to a point that is within the range of the standard curve when quantification by qPCR is performed. We dilute libraries to 10 pM, based upon their Bioanalyzer concentration, and quantify these by qPCR using 1, 10, and 100 pM standards. Standards should be made freshly if possible but can be reused if stored frozen in low-bind tubes (they should be discarded after ten freeze-thaw cycles and remade from a stock library at high concentration that has been stored frozen in a low-bind tube). Failure to store standards carefully results in degradation of standards and the qPCR quantification giving falsely high measurements. This leads to a lower cluster density than anticipated.

Cluster amplification QC

Too high a cluster density will result in a lower yield of purity-filtered data, because clusters will overlap to a greater extent. Conversely, too low a cluster density can result in a lower yield of data, and additional sequencing lanes being required. Figure 18.2.3 shows the range of cluster densities that will yield a good quantity of PF clusters, whereas outside of this range, decreased yields are inevitable.

Troubleshooting

Some problems that may be encountered in carrying out the protocols described in this unit, along with their possible causes and solutions, are described in Table 18.2.1.

Anticipated Results

The Illumina library preparation protocols and kits, when used with the recommendations discussed here, result in a robust approach that enables the preparation of high-quality libraries of adapter ligated fragments, of the correct concentration and quality for sequencing. Libraries should have the desired range of insert sizes and be free of adapter dimers, and should be at a concentration in the nanomolar range, that will allow for preparation of multiple flowcells for sequencing, with cluster densities of ∼160,000 clusters per tile for a GAII flowcell.

Time Considerations

The Illumina library preparation protocol can be completed within 1 day if processing 1 sample, and 2 days if processing multiple (2 to 8) samples. The protocol can be stopped after any column cleanup step, and samples can be stores in low-bind tubes at −20°C until required.

Agilent Bioanalyzer QC and qPCR take 0.5 days.

Cluster amplification, linearization, blocking, primer hybridization, and setting up the sequencing run can be performed in a single day, but if desired, flowcells can be stored after amplification. Once the sequencing primer has been hybridized, the sequencing run should be started within 4 hr.

Acknowledgements

This work was supported by the Wellcome Trust (grant number WT079643).

QIAGEN Exhibit 1033.0025

We are grateful to all members of Sequencing Technology Development, Illumina Library Construction, and Illumina Sequencing teams at the Sanger Institute.

Literature Cited

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M., and Gibbs, R.A. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4:903-905.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira, Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling, Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. and Smith, A.J. 2008. Accurate whole-human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59.

Hawkins, T.L., O'Connor-Morin, T., Roy, A., and Santillan, C. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res.* 22:4543-4544.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., and McCombie, W.R. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39:1522-1527.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945.

Mandel, M. and Marmur, J. 1968. Use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA. *Methods Enzymol.* 12:195-206.

Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24:133-141.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Maxam, A.M. and Gilbert, W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74:560-564.

Meyer, M., Briggs, A.W., Maricic, T., Hober, B., Hoffner, B., Krause, J., Weihmann, A., Paabo, S., and Hofreiter, M. 2008. From micrograms to picograms: Quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res.* 36:e5.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* 5:1005-1010.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J.C., and Markham, A.F. 1990. A novel, rapid method for the isolation of terminal sequences

QIAGEN Exhibit 1033.0026

from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res.* 18:2887-2890.

Ronaghi, M., Uhlen, M., and Nyren, P. 1998. A sequencing method based on real-time pyrophosphate. *Science* 281:363-365.

Sambrook, J., Fritsch, E., and Maniatis, T. 1989. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sanger, F. and Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94:441-448.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74:5463-5467.

Smith, D. and Malek, J. 2007. Asymmetrical adapters and methods of use thereof. USPTO Application no. 20070172839.

Surzycki, S. 2000. Basic Techniques in Molecular Biology. Springer-Verlag, Berlin.

QIAGEN Exhibit 1033.0027