Co-optimization of Probes and Polymerases to Drive the Evolution of Targeted Sequencing

Daniel Burgess[1], Maryke Appel[2], Erin Heinzen[3], Brian Krueger[3], Ryan Bannen[1], Michael Brockman[1], John Foskett[2], Bronwen Miller[2], Eric van der Walt[2], Dawn Green[1], Courtney Skalitzky[1], Jennifer Wendt[1]

[1]Roche NimbleGen, Madison, Wisconsin, USA, [2]KAPA Biosystems, Woburn, Massachusetts, USA, [3]Duke University School of Medicine, Durham, North Carolina, USA

Abstract

The evolution of targeted enrichment methods for high-throughput sequencing applications has focused optimization efforts onto a small number of persistent technical impediments to increased throughput and performance. Primarily, these include inefficiencies in library construction, automation-unfriendly workflows, and the well-documented tendency of many DNA polymerases to introduce strong biases against AT- and GC-rich targets. By combining enzymes specifically tailored for high performance NGS applications with innovative oligonucleotide probe design, we developed protocols for improved library construction and targeted enrichment. The result is a workflow that retains input sample complexity, minimizes amplification biases and artifacts such as PCR duplicates and chimeric library inserts, is compatible with manual or high-throughput workflows, and delivers unparalleled sensitivity for variant discovery over a wider range of targets throughout the genome than had previously been demonstrated.

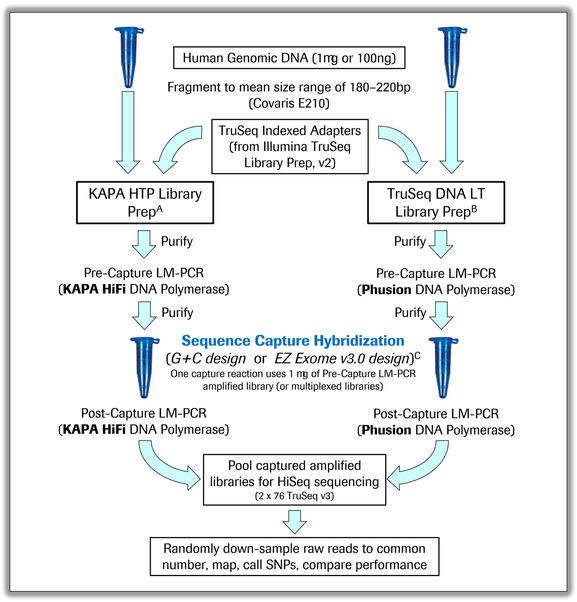
We applied the new reagent and protocol combinations to a series of enrichment targets, including human exomes and panels of genomic targets specifically designed to challenge capture performance with (A+T)-rich and (G+C)-rich targets. Sequencing of enriched material was performed on the Illumina HiSeq instrument. The results presented here extend the capabilities of targeted sequence enrichment, a method that has already transformed the work of genome analysis, and will enable the future discovery of more variation, in more regions of the genome, in more samples, and in less time.

Experimental Design

The currently recommended protocol for NimbleGen Sequence Capture (SeqCap EZ) prior to Illumina sequencing designates the input to be shotgun libraries constructed from 1mg of gDNA using the TruSeq DNA Sample Prep Kit (Illumina, Inc.), with Phusion polymerase (Thermo Fisher) used for all pre- and post-hybridization amplification steps. This particular combination of probes, library preparation, polymerase and processes has demonstrated superior performance over competing methodologies and products in every independent, peer-reviewed comparison published to date.[1-5] Nevertheless, a combination of anecdotes and hard evidence from the research community suggested there were still opportunities for improvement in both workflow and performance. To address this, we evaluated the following Sequence Capture workflow/protocol modifications:

1) **Use of the KAPA HTP Library Preparation Kit** (Kapa Biosystems; KK8234) in place of the TruSeq DNA LT Sample Prep Kit (Illumina; FC-121-2001).
   - § The KAPA kit was specifically designed for high-throughput workflows and automated liquid handling systems with convenient reagent volumes and kit formatting.
   - § Efficient and automation-friendly reaction cleanups are facilitated through a "with-bead" strategy[6] which reduces sample loss during library construction and results in significantly more correctly adapter-ligated fragments, greater retention of initial sample complexity, and potentially lower PCR duplicate rates after library sequencing.
2) **Use of KAPA HiFi HotStart DNA Polymerase** (Kapa Biosystems; KK2611), with optimized reaction protocols, in place of Phusion polymerase (Thermo Fisher; F-530).
   - § Thermostable DNA polymerases used in DNA sequencing have varying amplification efficiencies of (G+C)-rich or (A+T)-rich fragments, which can lead to significant locus "drop-out" and non-uniform sequence coverage over these regions.[7]
   - § KAPA HiFi polymerase was engineered for high-fidelity, high processivity and efficient amplification of targets across a broad range of (G+C) content, and has been demonstrated to outperform Phusion polymerase for uniform amplification of DNA fragments from (G+C)-rich, (A+T)-rich, or complex genomes.[8,9]

OVERVIEW



[A] KAPA HTP Library Preparation protocol for Illumina platforms (KR0426 – v2.12); [B] TruSeq DNA Sample Preparation Guide (Part # 15026486 Rev. C, July 2012); [C] Roche NimbleGen SeqCap EZ SR User Guide (v3.0 and v4.0)

SeqCap Probe Designs

SeqCap EZ Human Exome Library v3.0 (capture target = **64 Mb**)
http://www.nimblegen.com/products/seqcap/ez/v3/index.html

Human EZ G+C design (capture target = **1.87 Mb**)

We tiled through the hg19 genome at 1000 bp intervals. The first 250 bases of each 1000 bp window were evaluated as potential targets. Percent (G+C) was calculated for each 250mer and these were sorted into bins by (G+C) content. The 250mers in each of 13 bins had (G+C) content within a 5% range, except those within the bins for GC% ≤17.5% or >82.5% (Table 3).

We designed SeqCap EZ probes for all 250 bp targets within each bin and selected a single unique probe for each 250mer (average probe length = ~74bp). We then randomly selected 500 unique SeqCap EZ probes representing 500 different 250 bp target sequences from each bin (*Note: only 470 unique SeqCap EZ probes were available in the >82.5% bin*). These probes were manufactured into SeqCap EZ pools and used in capture experiments.

Optimization of Amplification Conditions for Kapa Sequencing Libraries and HiFi Polymerase

Pre-capture LM-PCR

1) Prepare Pre-Capture LM-PCR Master Mix (see Table 1 below; TS-PCR Oligo 1 and 2 are described in the *SeqCap EZ SR User Guide, v3.0 and v4.0*):

Table 1. Pre-Capture LM-PCR Master Mix

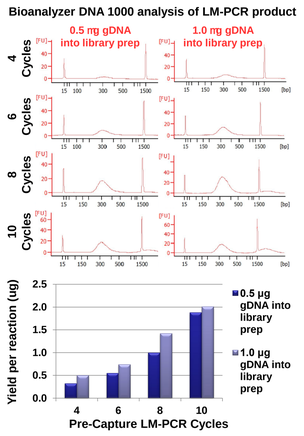
Pre-Capture LM-PCR Master Mix	One Rxn Per Capture
Kapa HiFi HS ReadyMix (2x)	25 µl
5 µM* TS-PCR Oligo 1	2.5 µl
5 µM* TS-PCR Oligo 2	2.5 µl
Total	**30 µl**

* *Increasing primer concentrations to 10 µM is advised for the amplification of libraries constructed with >500ng input gDNA, to reduce primer depletion artifacts*

2) Add 20 µl of sample library to the PCR tube containing the 30 µl LM-PCR Master Mix. Mix by pipetting up and down 5 times. Do not vortex.

3) Place the PCR tube in the thermocycler and amplify the sample library using the following Pre-Capture LM-PCR program:

Step 1: 45 seconds @ 98°C
Step 2: 15 seconds @ 98°C
Step 3: 30 seconds @ 60°C
Step 4: 30 seconds @ 72°C
Step 5: Go to Step 2, repeat **N** times
Step 6: 1 minute @ 72°C
Step 7: Hold @ 4°C



Post-capture LM-PCR

1) Prepare Post-Capture LM-PCR Master Mix (see Table 2 below):

Table 2. Post-Capture LM-PCR Master

Post-Capture LM-PCR Master Mix	Two Rxn Per Capture
Kapa HiFi HS ReadyMix (2x)	50 µl
5 µM* TS-PCR Oligo 1	5 µl
5 µM* TS-PCR Oligo 2	5 µl
Total	**60µl**

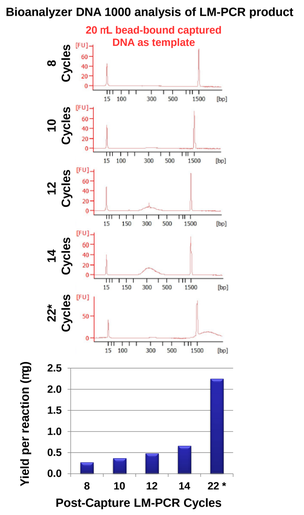
* *See comment below Table 1 above*

2) Pipette 30 µl of Post-Capture LM-PCR Master Mix into each of two reaction tubes.

3) Vortex the bead-bound captured DNA sample (streptavidin-biotin-probe-target complex) to ensure a homogenous mixture of beads.

4) Add 20 µl of bead-bound captured DNA sample to the PCR tube containing the 30 µl LM-PCR Master Mix. Mix well by pipetting up and down 5 times. Place the PCR tube in the thermocycler and amplify the sample library using the following Post-Capture LM-PCR program:

Step 1: 45 seconds @ 98°C
Step 2: 15 seconds @ 98°C
Step 3: 30 seconds @ 60°C
Step 4: 30 seconds @ 72°C
Step 5: Go to Step 2, repeat **N** times
Step 6: 1 minute @ 72°C
Step 7: Hold @ 4°C



***NOTE:** Intentional overamplification (22 cycles, see at right) was done to demonstrate a common artifact of LM-PCR over-amplification, where depletion of primers and/or dNTPs leads to arrest of complementary strand synthesis and non-covalent concatemerization of library fragments through the annealing of complementary adapter regions of library fragments with different inserts (i.e. "daisy-chaining"). The high molecular weight-appearing material is not an indication of a failed experiment and is competent to be sequenced.*

Experiment 1: Human EZ G+C Design Capture

Overview

To evaluate the relative performance of Sequence Capture protocols utilizing the TruSeq/Phusion versus the Kapa/Kapa Library Prep/Polymerase combination, we designed SeqCap EZ probes targeting a series of human genomic intervals within well-defined ranges of (G+C) content (Table 3).

Replicate captures were carried out using gDNA libraries with different adapter indexes, and the captured amplified libraries pooled for sequencing to minimize any effects of lane variability on read quality. The libraries were constructed from HapMap DNA sample NA12762 and amplified according to the protocols and conditions described in Panel 2 and Panel 4.

Raw reads were randomly down-sampled to 20M (10M read pairs) after demultiplexing, but before mapping and duplicate removal, to facilitate accurate comparisons of yield-dependent metrics. Mapping and SNP calling were performed using the SOAPv2 analysis package. The data from replicate captures were similar, so only one set of data is presented here.

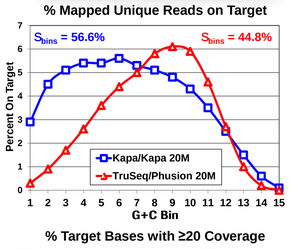
Table 3. Distribution of capture targets by (G+C) content in the SeqCap EZ G+C design

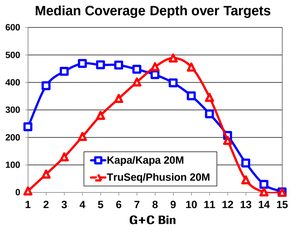
Bin	# of 250 bp targets in each Bin	%(G+C) content of 250 bp capture targets in each Bin
1	500	<17.5%
2	500	17.5 to 22.5%
3	500	22.5 to 27.5%
4	500	27.5 to 32.5%
5	500	32.5 to 37.5%
6	500	37.5 to 42.5%
7	500	42.5 to 47.5%
8	500	47.5 to 52.5%
9	500	52.5 to 57.5%
10	500	57.5 to 62.5%
11	500	62.5 to 67.5%
12	500	67.5 to 72.5%
13	500	72.5 to 77.5%
14	500	77.5 to 82.5%
15	470	>82.5

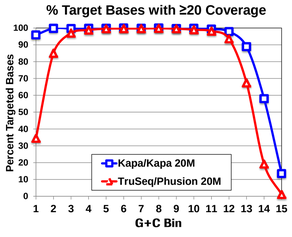
Results

Analysis of the sequence data indicated a dramatic advantage in the relative performance over both high (G+C) and low (G+C) targets for captures using libraries prepared with the Kapa HTP Library prep kit and amplified with the Kapa HiFi Polymerase. In the graphs below and right, the bin numbers on the x-axis correspond to the bin numbers in Table 3.

The Kapa protocol generated a greater percentage of mapped unique reads on-target, median coverage depth over targets, and percent targeted bases with a coverage depth ≥20 reads, for targets at both the high and low ends of the (G+C) spectrum. The performance gap was particularly notable among the lowest %(G+C) targets, where the Kapa protocol provided a 2.8-fold advantage in the fraction of bases with 20x minimum coverage depth among targets with <17.5% mean (G+C) content.







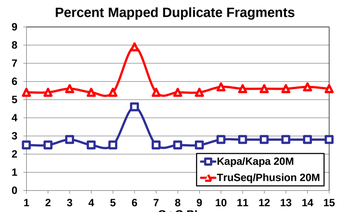
G+C Design Capture: Duplicate Rate and SNPs

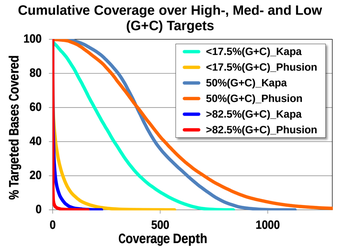
Results

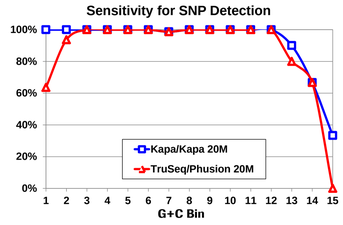
Increased rates of duplicate fragment sequencing reduce the efficiency of Sequence Capture experiments. Although this phenomenon can be caused by over-sequencing of a good quality capture product, particularly in the case of small capture targets, it can also be observed when correct amounts of sequencing are applied to a capture product with reduced complexity (other than that intended by the capture experiment). Reduced complexity can lead to allelic bias and reduced sensitivity for SNP detection in a capture experiment. We observed a rate of fragment duplication in captures from TruSeq libraries amplified with Phusion polymerase that was double that of those from the Kapa libraries amplified with Kapa HiFi polymerase (see graph at right). Although the difference in duplicate rate was constant over the entire range of (G+C) content, it is unclear if this phenomenon is more likely related to differences in polymerase performance or differences in library construction efficiency.

The cumulative coverage plot on the right provides additional perspective on the superior sequence coverage over the highest and lowest %(G+C) targets obtained using the Kapa-optimized capture protocol. It is interesting to note that this protocol also appears to provide higher coverage uniformity over more typical genomic regions, as indicated by the steeper inflection of the Kapa distribution curve relative to the Phusion for the 50% (G+C) target bin.

Distribution of coverage depth is an important metric for the evaluation of sequence capture efficiency, but it is only truly useful to the extent that it accurately predicts sensitivity for variant detection, which is the ultimate goal of the capture technology. We called SNPs from the G+C target capture experiments and compared these against previously genotyped (HapMap) SNPs from the same coordinates in DNA sample NA12762. The graph at right supports the correlation between coverage over (A+T) and (G+C) rich targets and an improved ability to detect SNPs in these regions through the use of Sequence Capture protocols optimized for use with Kapa library prep and amplification reagents and processes.







Experiment 2: SeqCap EZ Exome v3.0

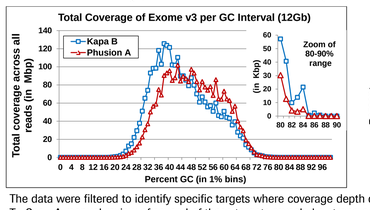
Overview

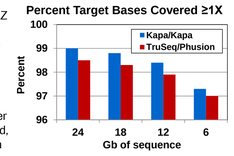
To examine the performance of KAPA HTP Library Prep and HiFi HotStart Polymerase in Sequence Capture from a target of greater biological relevance, we used NimbleGen SeqCap EZ Human Exome v3 probes to capture from human genomic DNA libraries prepared using the Kapa kit with HiFi Polymerase or the TruSeq kit with Phusion polymerase. The libraries were constructed from HapMap DNA sample NA12762 and amplified according to the protocols and conditions described in Panel 2 and Panel 4. Raw reads (2x76bp sequencing) were randomly down-sampled to 160M, 120M, 80M or 40M to facilitate accurate comparisons of yield-dependent metrics. Mapping and SNP calling were performed using the SOAPv2 analysis package.

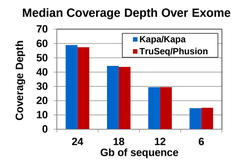
Results

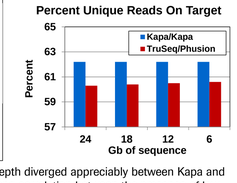
Analysis of the results indicated a roughly similar performance for EZ Exome v3 captures using either the Kapa or TruSeq protocols, for percent target bases covered ≥1X, median coverage depth over the exome, and percent unique reads on target (see graphs at right). The notable exception was percent mapped duplicate fragments, which was 2 to 3-fold greater for the TruSeq captures, consistent with the G+C design capture results (data not shown). However, because high (A+T) and (G+C) targets will comprise a much smaller fraction of the exome design compared to the G+C design we tested, the small performance advantage we observed with use of the Kapa protocol (*i.e.* 2% greater percent unique reads on-target, 0.5% more target bases covered ≥1X) could be very meaningful in the limited context of those small numbers of high (A+T) and (G+C) targets.

To examine this possibility more closely, we compared total coverage over Exome v3 by %(G+C) content for 12Gb sequence (80M reads). We mapped the data with BWA 0.6.2 and then used GATK v1.6 to calculate total coverage (Mbp) for all reads in the sequencing run and this was binned according to %(G+C) content of target regions. The results are shown below. The inset shows a zoom of the 80-90% range, with the Y-axis at a different scale (Kbp).

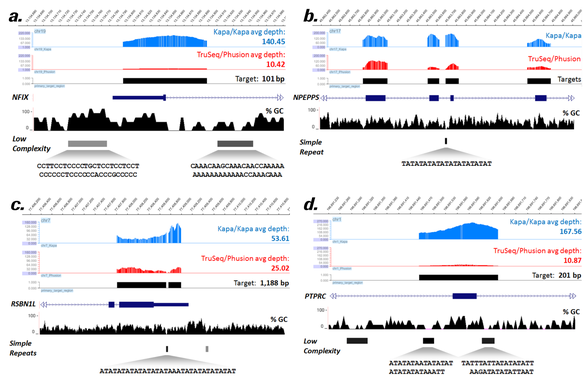








The data were filtered to identify specific targets where coverage depth diverged appreciably between Kapa and TruSeq. A manual review of several of these targets revealed a strong correlation between the presence of low %(G+C) simple sequence repeats or low-complexity sequence and poor coverage depth in the TruSeq/Phusion data relative to the Kapa/Kapa data. High %(G+C) targets were less frequently observed. Representative examples are shown below (*panels a-d*).



Experiment 3: Multiplex SeqCap EZ Exome v3.0

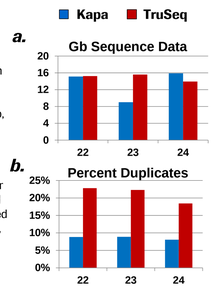
Overview

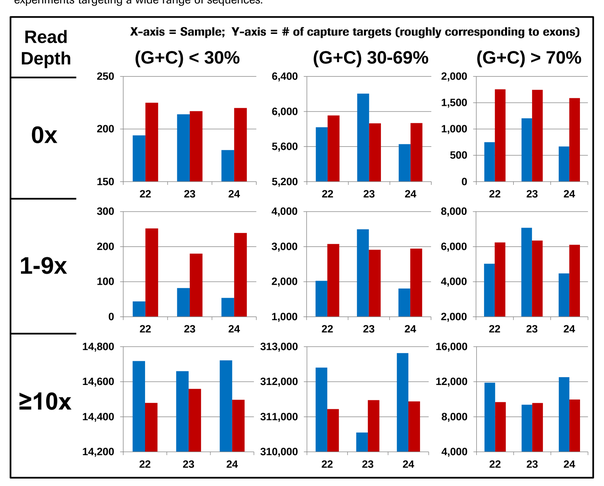
The performance of the KAPA HTP "with-bead" Library Prep with HiFi HotStart Polymerase was compared to the performance of the TruSeq Library Prep kit with Phusion polymerase for EZ Exome v3.0 capture from multiplexed (3-plex) human DNA libraries upstream of Illumina 2x100 HiSeq sequencing. The same three samples (22, 23, 24) were used for both workflows in the study. 1mg of gDNA was used in each library prep, with TruSeq indexed adapters. Pre-capture amplification was 7 cycles (Kapa HiFi) or 8 cycles (Phusion). The threesamples prepared using the same method were pooled in equal amounts, and 1mg of total pooled amplified sample library was used in the capture hybridizations as described in the Roche NimbleGen SeqCap EZ SR User Guide (v4.0). For post-capture amplification, the yields were 1.93mg (Kapa, 14 cycles) and 0.63mg (TruSeq/Phusion, 18 cycles). Sequencing data were demultiplexed and processed using CASAVA v1.8 prior to filtering and quality trimming, and then alignment to the Human Reference Genome (NCBI Build 37) using the Burrows-Wheeler Alignment Tool (BWA) (version 0.5.10).

Results

Basic mapping and coverage statistics are shown at right, with primary comparisons performed between the same DNA samples processed with the two different library preps and polymerases (*a-d*). Similar amounts of data were generated and compared for samples 22 and 24, but much less data was generated for the sample 23 Kapa library relative to the TruSeq library, likely due to inaccurate quantification or pipetting during pooling prior to capture (*a*). This discrepancy would be expected to favor the sample 23 TruSeq library in subsequent coverage-dependent comparisons to a similar degree. Percent duplicates were significantly higher for the TruSeq data compared to the Kapa data (*b*).

For this study, we focused our analysis on the distribution of coverage over low (<30%), medium (30-69%), and high (>70%) G+C targets in the exome (see panel below). Consistent with the previously described experiments, the data obtained through use of the Kapa-optimized protocol covered a greater number of targets in the low and high G+C categories to a greater depth than the data from the TruSeq protocol. These performance improvements, obtained through use of the Kapa HTP library prep and HiFi polymerase, are expected to significantly increase the sensitivity for variant detection in Sequence Capture experiments targeting a wide range of sequences.





Observations & Conclusions

- Use of the KAPA HTP Library Preparation Kit with KAPA HiFi HotStart Polymerase provided superior performance in NimbleGen Sequence Capture experiments compared to use of the TruSeq DNA LT Sample Prep Kit with Phusion polymerase.
- Use of the KAPA HiFi Polymerase in Sequence Capture facilitated greater LM-PCR yields with fewer cycles, a result enhanced by improved ligation efficiency and lower DNA loss with the KAPA "with-bead" library prep, and was correlated with 30-50% lower duplicate fragment rates compared to TruSeq/Phusion based protocols.
- The "Kapa/Kapa" protocol modification combined with NimbleGen SeqCap EZ probes provided deeper coverage than the "TruSeq/Phusion" capture protocol over difficult-to-capture high (G+C) and (A+T) targets.
- Single isolated NimbleGen SeqCap EZ probes (avg length ~74-mer) are sufficient to capture DNA fragments over a very wide range of G+C content, indicating that much larger capture targets are feasible with a single 2.1M probe design than had previously been demonstrated.

References

1. Performance comparison of exome DNA sequencing technologies. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M. Nat Biotechnol. 2011 Sep 25;29(10):908-14.
2. A comparative analysis of exome capture. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. Genome Biol. 2011 Sep 29;12(9):R97.
3. Comparison of solution-based exome capture methods for next generation sequencing. Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, Joensuu H, Raivio T, Suomalainen A, Saarela J. Genome Biol. 2011 Sep 28;12(9):R94.
4. Comprehensive comparison of three commercial human whole-exome capture platforms. Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, Wang J, Wang J, Yang H, Zhang X. Genome Biol. 2011 Sep 28;12(9):R95.
5. Assessing the enrichment performance in targeted resequencing experiments. Frommolt P, Abdallah AT, Altmüller J, Motameny S, Thiele H, Becker C, Stemshorn K, Fischer M, Freilinger T, Nürnberg P. Hum Mutat. 2012 Apr;33(4):635-41.
6. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, Young G, Fennell TJ, Allen A, Ambrogio L, Berlin AM, Blumenstiel B, Cibulskis K, Friedrich D, Johnson R, Juhn F, Reilly B, Shammas R, Stalker J, Sykes SM, Thompson J, Walsh J, Zimmer A, Zwirko Z, Gabriel S, Nicol R, Nusbaum C. Genome Biol. 2011;12(1):R1.
7. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. BMC Genomics. 2012 Jul 24;13:341.
8. Optimal enzymes for amplifying sequencing libraries. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, Oyola SO. Nat Methods. 2011 Dec 28;9(1):10-1.
9. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. BMC Genomics. 2012 Jan 3;13:1.