Master's Thesis

Implementation and Comparative Assessment of Diagnostic Cancer Gene Panels in the Molecular Pathology Laboratory

University of Luxembourg

Faculty of Science, Communication and Technology

Master in Integrated Systems Biology

by

Ben Flies

(010081174D)

Abstract

Contents

List of Abbreviations

NGS	Next Generation Sequencing
LNS	Laboratoire National de Sante
SGMB	Service of Genetics and Molecular Biology TST15
Illumina TruSight Tumor 15	

List of Figures

List of Tables

1 Introduction

Cancer represents a huge burden for health care systems worldwide and one of the leading death causes. In 2012, there were an estimated 14.1 million new cancer cases with estimated 8.2 million cancer deaths [3]. Lung cancer is the most common cancer, both in terms of new cases (1.8 million) and deaths (1.6 million). Breast cancer is the second most common cancer (1.7 million cases) but only ranks 5th as cause of death (522,000 deaths). Colorectal cancer (1.4 million cases; 694,000 deaths), prostate cancer (1.1 million cases; 307,000 deaths), stomach cancer (951,000 cases; 723,000 deaths) and liver cancer (782,000 cases; 723,000 deaths) are following.

Scientific discoveries in the last decade have had an enormous impact on our understanding of the underlying causes of cancer. The development of omics techniques, in combination with enhanced computational power, has lead to an explosion of biological data. It has become clear that cancer is an incredibly complex malignancy. The research community is trying to interprete this vast amount of data with the goal to get a deeper understanding of cancer and to cure it eventually. In recent years, several drugs have been approved that target proteins needed for cancer development, proliferation or metastasis. Molecular testing is employed to check whether these targeted drugs would be of benefit. In that regard, Next-Generation Sequencing (NGS) is an interesting method to gain deep insights into the genetic information of a tumor and to guide personalized therapy.

1.1 Cancer Genetics

Cancerogenesis is considered to be caused by an imbalance between the occurrence of mutations and cell-cycle control mechanisms, thereby leading to cardinal features of cancer: genomic instability and modifications. These alterations are caused either by inherited mutations or are acquired during cancerogenesis. Critical alterations include single nucleotide variations (SNVs), insertions and deletions of one or multiple nucleotides (INDELS), copy number variations (CNVs) and rearrangements.

Genomic instability can affect the genetic information at the level of nucleotides, microsatellites, whole genes or chromosomes. For instance, chromosome rearrangements, chromosome number alterations, loss of heterozygosity (LOH) or gene amplification contributee to chromosomal instability (CIN), which occurs in 50–85% of colorectal cancers (CRCs).

Genomic modifications in tumors lead to the inactivation of tumor-suppressor genes and / or the activation of oncogene pathways. The normal activity of tumor-suppressor genes acts to limit cancer development and proliferation by, for instance, controlling the cell cycle or cell motility. This protection is lost in many cancers through several mechanisms that include SNVs, deletions, loss of one allele with mutation on the other and promoter methylation. Genes affected by these modifications include TP53, APC and TGF–. p53 and its coding gene TP53 are well-known to be master regulators, which are misregulated in many cancer types. The inactivation of TP53 by LOH and mutations is a crucial step in many CRCs.

Oncogenes encode for signalling molecules, cell-cycle regulators, growth factors and their receptors. Epigenetic / transcriptional modifications, copy number amplifications or modifications that impair the normal function of the protein lead to overexpressed or overactive gene products, giving the cell, in which the alteration occurs, a proliferative advantage.

Driver mutations

1.2 Targeting the EGFR Pathway in Solid Tumors

EGFR Pathway

Targeted Drugs

Resistances

1.3 Targeted Sequencing

1.3.1 Target Enrichment Methods

1.3.2 Illumina Sequencing Chemistry

1.4 NGS Data Analysis

1.4.1 GATK Best Practices

1.5 Practical Implications in the Laboratory

1.6 Aims of the Thesis

2 Material and Methods

- How was the data analyzed ?
- Present econometric/statistical estimation method and give reasons why it is suitable to answer the given problem.
- Allows the reader to judge the validity of the study and its findings.
- Depending on the topic this section can also be split up into separate sections.

2.1 Library Preparation

2.1.1 Patients

Melanoma Non-Small Cell Lun Carcinoma (NSCLC) metastatic colorectal cancer (mCRC) Chronic lymphocytic leukemia (CLL) were extracted from blood and did not undergo FFPE treatment -¿ were used as some kind of good quality samples to see if there are really more C¿T variants in FFPE samples.

2.1.2 DNA Extraction, Quantification and Quality Control

DNA Extraction Kit from Qiagen

Quantification Qubit fluorometer, either High Sensitivity kit or Broad Range

Quality Control Illumina Infinium FFPE QC Assay kit

2.1.3 Agilent Haloplex ClearSeq Cancer

2.1.4 Illumina TruSight Tumor 15

2.2 Bioinformatic Analysis

2.2.1 Agilent SureCall

With alignment algorithms installed Windows System, 3.00GHz, 16GB RAM$ modifiable

2.2.2 Illumina BaseSpace TruSight Tumor 15 App

Cloud-based parameters not modifiable

2.2.3 Custom In-House Pipeline (Velona)

Linux System

2.2.4 Variant Calling Algorithms

Tested on Linux Ubuntu 14.04.4 LTS Ťrusty Tahrínstalled on VMware virtual box with of 12GB RAM

GATK HaplotypeCaller

VarScan 2

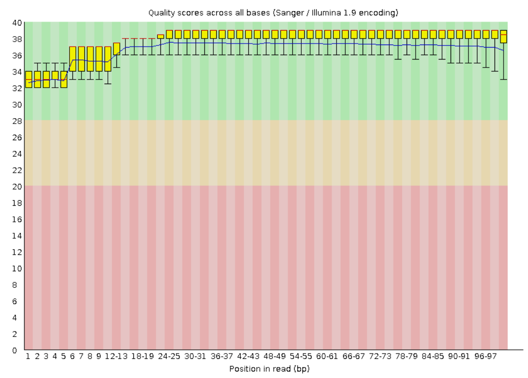
Mutect1.1.7 [2]

SomVarIUS

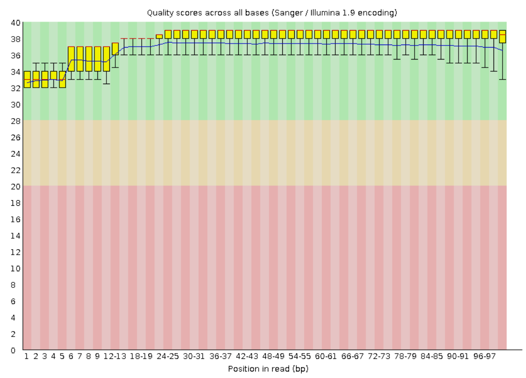
3 Results

3.1 Sample Preparation

Before pooling the adaptor-ligated and indexed sequencing libraries, the sucess of library preparation is validated using the Agilent Bioanalyzer instrument. 1a and 1b show representative electropherograms of a sample that has been processed using both kits. The expected DNA products should be detected at 175-600 bp for Haloplex CSC and 200-400 for TST15.



(a) Agilent Haloplex CSC (High Sensitivity DNA Chip)



(b) Illumina TST15 (MixA) (DNA 1000 DNA Chip)

Figure 1: Electropherograms of representative sequencing libraries prepared by Agilent Haloplex ClearSeq Cancer and Illumina TruSight Tumor 15. (*) represents the lower marker, (**) represents the upper marker

Using the blablabla software, the concentration, molarity and total peak area (TPA) of the expected sequencing libraries were calculated.

Maybe: relationship between dCt and TPA?

Maybe: the enrichment of one or the other kit is more affected by bad quality samples

3.2 NGS Data Quality

Sequencing run parameters were calculated by the Illumina Sequencing Viewer software. Table XXX shows the averaged run parameters of runs with Haloplex CSC and TST15 sample preparation.

TST15 has a higher cluster density and therefore a higher total yield, but has lower reads passing a phred-score threshold of Q30 than Haloplex. This is due to the different chemistries used. TST15 uses v3 chemistry, while Haloplex uses v2. v2 generally has lower cluster density and output, but

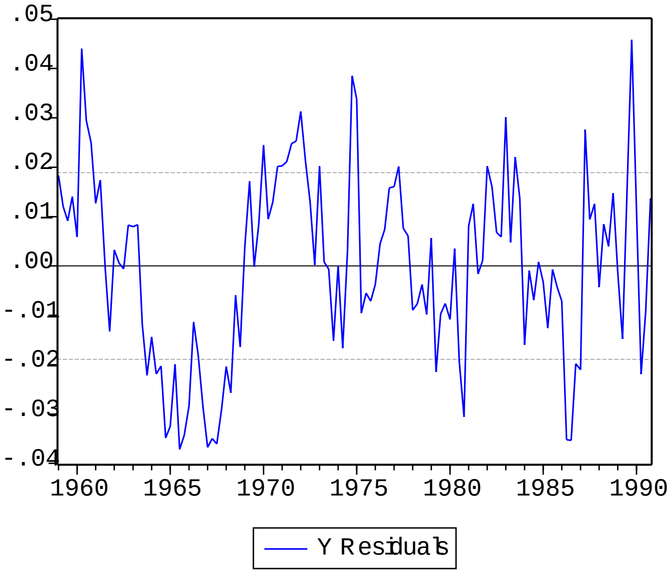
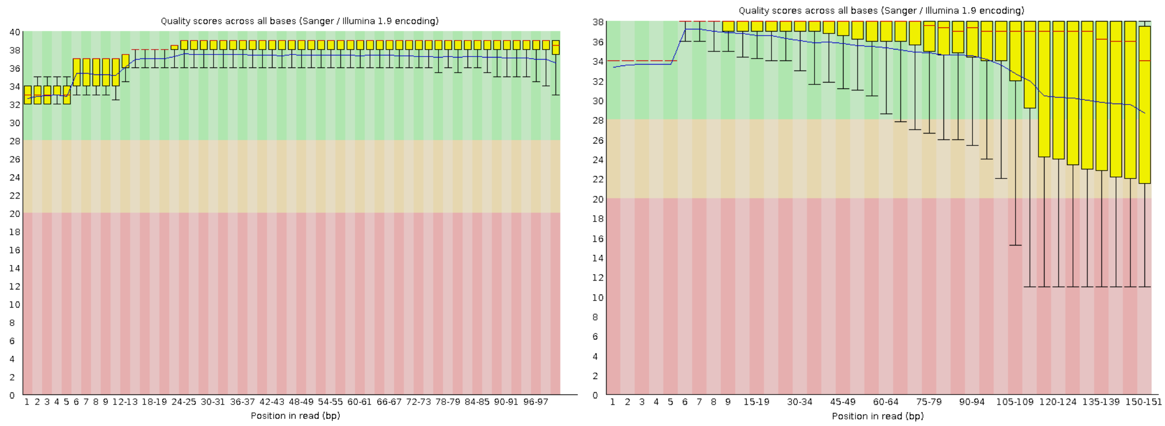


Figure 2: Scatter plot of the corrected peak area (X axis) of the regions corresponding to the sequencing libraries defined in the blabla software and the dCt (Y axis). Agilent Haloplex ClearSeq Cancer data are represented as blue dots, Illumina TruSight Tumor 15 data are represented as red dots.

Table 1: Comparison of Run Parameters (Averaged) of Sequencing Runs with Haloplex CSC & TST15 Sample Preparation

Parameter	Halo CSC	TST15
Yield total (Gb)	3.7	7.37
% >Q30	93.8	82.355
Cluster Density PF (k/mm2)	1084	1180
Cluster Density PF (%)	85.95	79.95



(a) Agilent Haloplex ClearSeq Cancer

(b) Illumina TruSight Tumor 15

Figure 3: Comparison of coverage distributions per amplicon as reported by FastQC

therefore better quality.

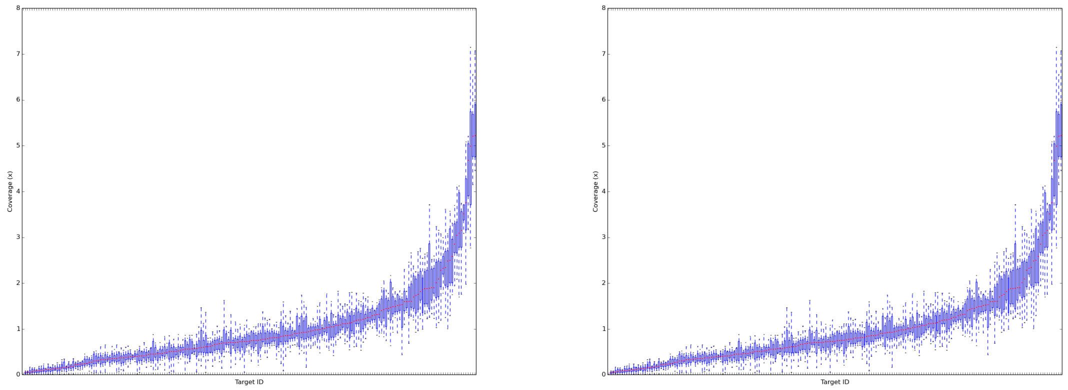
Table XXX shows the boxplot representations of the read qualities per position of two representative FASTQ files as reported by FastQC. Both workflows yield high quality data, yet Haloplex CSC data have more narrow distributions and are of higher quality. This is in direct relationship with the sequencing chemistry kit used.

Table 2: Blablabla

Parameter	Haloplex SureCall A	Haloplex SureCall B	Haloplex Velona	TST15 BaseSpace	TST15 Velona
% mapped	–	–	–	62.6	–
% paired	–	–	–	58.7	–
% singletons	–	–	–	3.8	–

The Samtools Flagstat command was used to determine some basic BAM statistics of BAM files of samples prepared with the respective library preparations and processed with the mentioned bioinformatic pipelines. **??** shows the averaged result of these statistics.

Considering the recommended pipelines, Haloplex CSC data, analyzed with Agilent's SureCall software, has a higher percentage (91%) of mapped reads when compared to Illumina's BaseSpace TruSight Tumor 15 App (62.9%). Data analysis with the recommended SureCall design includes a steps where mates are fixed, but they are not stitched together. Therefore no reads are considered as being paired. The TST15 app in contrast includes a read stitching step and 58% are considered as properly paired. This means that of the 62.6% of mapped reads, 4.2% are not properly paired.



(a) Agilent Haloplex ClearSeq Cancer

(b) Illumina TruSight Tumor 15

Figure 4: Comparison of Coverage Distributions per Amplicon

3.8% of reads processed with the TST15 online App are considered to be singletons, whereas only 1.8% of reads processed with the SureCall software are considered as singletons.

3.3 Coverage Analysis

Coverage Distribution TST15 vs Haloplex

Coverage Distribution per Patient (check if correlation IQR with dCt)

Coverage Distribution per Amplicon (check if some have always lower coverage, check if some failed)

Failed Amplicon Counter

Table 3: Blablabla

Amplicon	1x	50x	250x	500x	1000x
ATM_14	12	0	0	0	0
MAP2K1_2	1	0	0	1	4
ATM_5	0	1	1	0	2
ATM_11	0	1	1	2	3
PTEN_1	0	1	1	0	2
KIT_6	0	1	0	1	2
FGFR3_1	0	0	1	2	2

Table 4: Blablabla

Amplicon	1x	50x	250x	500x	1000x
1METxxxE16TF031SR031	0	11	0	1	3
2KITxxxE09TF003SR003	0	5	6	0	0
2KITxxxE09TF003SR003	0	5	6	0	0
27qxxxxExxTF034SR034	0	0	2	4	5
17qxxxxExxTF018SR018	0	1	1	4	4
1KRASxxE04TF002SR002	0	1	1	3	6
1TP53xxE02TF034SR034	0	0	3	5	9
1EGFRxxE19TF032SR032	0	0	1	5	5
1EGFRxxE21TF035SR035	0	0	1	2	2
2TP53xxE02TF033SR033	0	0	0	1	3

Table XXX and table XXX show how often a given amplicon failed a given coverage threshold. The number of failed amplicons is low for Haloplex CSC as wel las for TST15. Most amplicons were amplified efficiently in most samples, which is also confirmed by figure XXX (amplicon distributions). Some amplicons however fail coverage thresholds in several samples.

- Amplicon ATM_14 in Haloplex CSC was never amplified. This amplicon is defined in the BED file but obviously is not part of the kit. (negative control?)
- Amplicons ATM_5, ATM_10, MAP2K1, PTEN_1, KIT_6 and FGFR3_1 in Haloplex CSC data failed a coverage threshold of 1000x in several samples
- In TST15 data, more amplicons fail the respective thresholds. Several amplicons in genes MET, KIT, TP53, KRAS and EGFR fail the 1000x coverage threshold, and often even the required threshold of 500x, which is required by the TruSight Tumor 15 App.

The fact that several amplicons in the genes EGFR and KRAS in TST15 data repetitively fail the required coverage thresholds is problematic. This is especially the case for amplicon 1EGFRxxE21TF035SR035 as it includes the well-known EGFR L858R variant, which confers increases sensitivity for EGFR tyrosine kinase inhibitors.

Fragmentation ¡-¿ Coverage?

(GATK CallableLoci) (GATK CountLoci???) (GATK FindCoveredIntervals)

On-off target; Enrichment Efficiency TST15 vs Haloplex

Coverage across genome, check where there is coverage

Strandedness?

GATK DepthOfCoverage???

3.4 Variant Calling Algorithm Comparison

3.4.1 Detection of Known Single Nucleotide Variants and Deletions

Table XXX shows known variants in the analyzed samples and the variant frequency reported by the recommended pipelines. There is a high concordance between the results of both kits and previously known variants.

TST15 could not detect EGFR del19 in patient F due to low coverage. The corresponding region was inspected in IGV and the deletion was present. The amplicon was not amplified efficiently and has only a coverage of 167x. The TST15 BaseSpace App however applies a coverage threshold at 500x. Variants with a lower depth are not reported. The same sample was sequenced twice and the deletion was never detected. This is probably due to the fragmentation induced by the FFPE fixation.

Haloplex CSC did not find KRAS p.Gly12Val in patient G, also due to low coverage for this amplicon. The corresponding region was inspected in IGV: the region has a coverage of only 180x. Only one read showed the expected C¿A variant. Sample G is also the sample with the worst dCt (2.85). The hig fragmentation in this sample is probably responsible for the bad amplification. The same sample will be re-sequenced in a later run to check is the problem is related to the fragmentation of the sample or if library preparation was bad.

Additional previously unknown variants were found (TODO: put a table into the appendix)

3.4.2 Sensitivity Analysis

BRAF Mut and WT samples from Horizon were analyzed. The BRAF Mut sample was sequenced purely, as well as in a 1/3 and 2/3 dilution with the BRAF WT sample.

The observed variant frequencies as detected by the TST15 BaseSpace App are in line with the expected variant frequencies. D816V variant in cKIT could not be observed as the position of this

Table 5: Blablabla

Sample	Context	Tissue	Known variant	Halo CSC Freq (%)	TST15 Freq (%)
A	NSCLC	FFPE	EGFR L858R	24.7	21.4
B	NSCLC	FFPE	EGFR L858R	24.3	13.2
C	NSCLC	FFPE	EGFR L858R	32.7	29.8
D	Melanoma	FFPE	BRAF V600E	44.4	47.6
E	Melanoma	FFPE	BRAF V600E	18.4	21.4
F	NSCLC	FFPE	EGFR del19	not found	50
G	mCRC	FFPE	KRAS CD 12_13	p.Gly12Val (3.4)	not found
H	mCRC	FFPE	KRAS CD 12_13	p.Gly12Asp (37.4)	G12D (31.4)
I	mCRC	FFPE	KRAS CD 12_13	p.Gly13Asp (9.7)	G13D (8.7)
J	mCRC	FFPE	NRAS p.Gly12Asp	25.9	27.2
K	Melanoma	FFPE	BRAF V600E	66.2	59.5
L	mCRC	FFPE	NRAS p.Gly13Val	6.6	5
M	mCRC	FFPE	KRAS and NRAS WT	WT	WT
N	mCRC	FFPE	KRAS and NRAS WT	WT	WT
O	Melanoma	FFPE	BRAF V600E	37.4	

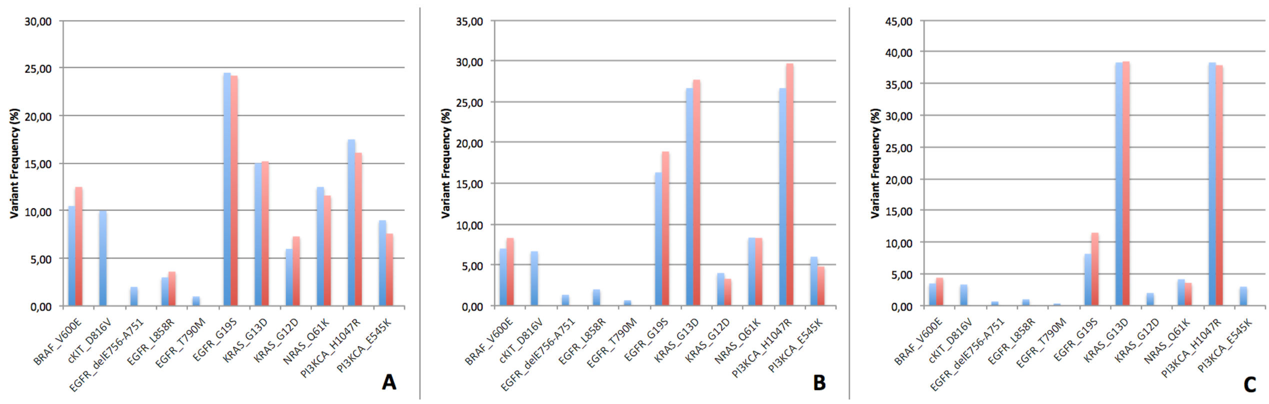


Figure 5: Blablabla

Table 6: Blablabla

Gene	Variant	Exp 100%	Obs Halo CSC	Obs TST 15	Exp at 66%	Obs Halo CSC	Obs TST15	Exp at 33%	Obs Halo CSC	Obs TST15
BRAF	V600E	10.5	–	12.5	7	–	8.3	3.5	–	4.4
cKIT	D816V	10	–	–	6.67	–	–	3.33	–	–
EGFR	delE756-A751	2	–	–	1.33	–	–	0.67	–	–
EGFR	L858R	3	–	3.6	2	–	–	1	–	–
EGFR	T790M	1	–	–	0.67	–	–	0.33	–	–
EGFR	G719S	24.5	–	24.2	16.33	–	18.9	8.17	–	11.5
KRAS	G13D	15	–	15.2	26.67	–	27.7	38.33	–	38.5
KRAS	G12D	6	–	7.3	4	–	3.3	2	–	–
NRAS	Q61K	12.5	–	11.6	8.33	–	8.3	4.17	–	3.6
PIK3CA	H1047R	17.5	–	16.1	26.67	–	29.7	38.33	–	37.9
PIK3CA	E545K	9	–	7.6	6	–	4.8	3	–	–

variant is not covered by the TST15 kit.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3986649/

TODO: do the same with Haloplex CSC

TODO: call variants with MuTect 1.1.7, VarScan 2, GATK HaplotypeCaller, SomVarIUS?, Freebayes?, Vardict??????? TODO: compare results

Among the variants detected by MuTect, 40–50 % are C¿T variants. This has been reported in several studies. TODO: do this for all samples, check if this is really statistically significant or only happened in a few samples

Tools that may be of use somehow: GATK SelectVariants; GATK VariantFiltration; GATK VariantEval; GATK ValidateVariants

4 Conclusions

References

[1] M. Berg, "EGFR and downstream genetic alterations in KRAS/BRAF and PI3K/AKT pathways in colorectal cancer – implictaions for targeted therapy," *Discovery Medicine*, vol. 14, no. 76, pp. 207–214, 2012.

[2] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Computational Biology*, 2013.

[3] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012," *International Journal of Cancer*, 2015.

[4] R. d. d. Direction de la Santé, Service des statistiques, "Statistiques des causes de décès pour l'année 2014," 2014.

[5] F. Mertes, A. E. Sharawy, S. Sauer, J. M. L. M. van Helvoort, P. van der Zaag, A. Franke, M. Nilsson, H. Lehrach, and A. J. Brookes, "Targeted enrichment of genomic dna regions for next-generation sequencing," *Briefings in Functional Genomics*, vol. 10, no. 6, pp. 374–386, 2011.