COMPUTATIONAL MODELING FOR IDENTIFICATION OF LOW-FREQUENCY SINGLE NUCLEOTIDE VARIANTS

Yangyang Hao

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Medical and Molecular Genetics,
Indiana University

February 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Yunlong Liu, PhD, Chair

Doctoral Committee

Howard J. Edenberg, PhD

November 16, 2015

Lang Li, PhD

Harikrishna Nakshatri, PhD

ACKNOWLEDGEMENTS

I would first and foremost like to thank my mentor, Dr. Yunlong Liu, for all the advice, support, motivation, and encouragement over the last five and a half years. Dr. Liu provided endless insight and ideas into helping me complete the thesis project and also many valuable training and collaboration opportunities to hone my skills as a young bioinformatics scientist throughout my time in his laboratory. He truly cares about the well-being of the students under his supervision, not only academically, but also skills and qualities that will benefit the students for a lifetime, such as effective communication, time management and a strong yet flexible mind facing problems, for which I will always be grateful.

I sincerely thank each of my committee members, Dr. Howard J Edenberg, Dr. Lang Li and Dr. Harikrishna Nakshatri, for sharing of their time and knowledge with me. Their constructive criticisms, insightful comments, and encouragements helped to shape my research for the better and I am truly appreciative. Particularly, I thank Dr. Edenberg for his great patience with my very poorly written drafts and his numerous valuable advices on the structure, logic flow and presentation of key points that can help deliver them better to the audiences. I thank Dr. Li for his insightful suggestions on data modeling that help to extend the horizon of my computational knowledge. I thank Dr. Nakshatri for offering biological insights that are relevant to my research and helping me put the computational problem into an important biomedical research and application perspective.

I thank the sequencing core in Center for Medical Genomics, for generating the sequencing data, especially Dr. Xiaoling Xuei, for her help on answering sequencing technology related questions. I thank Dr. Hongyu Gao from Center for Computational Biology and Bioinformatics for her support on using computational tools to process next-generation data.

I would also like to thank the faculty, staff, and students of the Indiana University School of Medicine Department of Medical and Molecular Genetics. I enjoyed every interaction with them and am grateful for their help and support. I would like to particularly thank Dr. Brittney-Shea Herbert for her super fast, informative replies to each of my email inquiry. I truly appreciate her kind reminders as well as support for everything needed to be a successful graduate student.

I owe a huge thank you to my family and friends for their love and support throughout this journey. I thank my parents for their never-ending encouragement and unwavering confidence in me. I thank my husband for being such a great person, my soul mate as well as the best team partner that I can ever imagine. I thank my dear friend Xue Wu, for encouraging me during the not-so-easy Ph.D. life, planning all fun travelling across the US, as well as her great baking skills that I will always want more. I wouldn’t be who I am today and where I am today without the constant support, encouragement, and inspiration of those I cherish most, for which I am eternally grateful.

Yangyang Hao

COMPUTATIONAL MODELING FOR IDENTIFICATION OF LOW-FREQUENCY SINGLE NUCLEOTIDE VARIANTS

Reliable detection of low-frequency single nucleotide variants (SNVs) carries great significance in many applications. In cancer genetics, the frequencies of somatic variants from tumor biopsies tend to be low due to contamination with normal tissue and tumor heterogeneity. Circulating tumor DNA monitoring also faces the challenge of detecting low-frequency variants due to the small percentage of tumor DNA in blood. Moreover, in population genetics, although pooled sequencing is cost-effective compared with individual sequencing, pooling dilutes the signals of variants from any individual. Detection of low frequency variants is difficult and can be cofounded by multiple sources of errors, especially next-generation sequencing artifacts. Existing methods are limited in sensitivity and mainly focus on frequencies around 5%; most fail to consider differential, context-specific sequencing artifacts. To face this challenge, we developed a computational and experimental framework, RareVar, to reliably identify low-frequency SNVs from high-throughput sequencing data. For optimized performance, RareVar utilized a supervised learning framework to model artifacts originated from different components of a specific sequencing pipeline. This is enabled by a customized, comprehensive benchmark data enriched with known low-frequency SNVs from the sequencing pipeline of interest. Genomic-context-specific sequencing error model was trained on the benchmark data to characterize the systematic sequencing artifacts, to derive the position-specific detection limit for sensitive low-frequency SNV detection. Further, a machine-learning algorithm utilized sequencing quality features to refine SNV candidates for higher specificity. RareVar outperformed existing approaches, especially

at 0.5% to 5% frequency. We further explored the influence of statistical modeling on position specific error modeling and showed zero-inflated negative binomial as the best-performed statistical distribution. When replicating analyses on an Illumina MiSeq benchmark dataset, our method seamlessly adapted to technologies with different biochemistries. RareVar enables sensitive detection of low-frequency SNVs across different sequencing platforms and will facilitate research and clinical applications such as pooled sequencing, cancer early detection, prognostic assessment, metastatic monitoring, and relapses or acquired resistance identification.

Yunlong Liu, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BIC	Bayesian information criterion
BWA	Burrows-Wheeler aligner
BWT	Burrows-Wheeler transform
CDCV	Common Disease-Common Variants
CDK4/6	Cyclin-dependent kinases 4 and 6
CDRV	Common Disease-Rare Variants
ctDNA	Circulating Tumor DNA
DAXX	Death Domain-Associated Protein 6
DCIS	Ductal Carcinoma In Situ
DNA	Deoxyribonucleic Acid
FFPE	Formalin-Fixed and Paraffin-Embedded
FOXO1	Forkhead Box Protein O1
FZR1	Fizzy/Cell Division Cycle 20 Related 1
GLM	Generalized Linear Model
IG	Information Gain
LCIS	Lobular Carcinoma In Situ
MAF	Minor Allele Frequency
MYH9	Myosin, Heavy Chain 9, Non-Muscle
NB	Negative Binomial
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
PD-GLM	Poisson Distribution Generalized Linear Model
PIK3R1	Phosphoinositide-3-Kinase, Regulatory Subunit 1
PSEM	Position Specific Sequencing Error Modeling

RNA	Ribonucleic Acid
SBS	Sequencing by Synthesis
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
UCSC	University of California-Santa Cruz
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

Chapter 1. Introduction and Literature Review

1.1 Importance of Low Frequency SNVs in Biomedical Research and Applications

Single nucleotide variants (SNVs) are the most common type of variation [1, 2], and are also currently the major data source to derive drug targets and biomarkers [3]. Thus identifying SNVs and studying the function of SNVs constitute key aspects in the realm of genomics and genetics. Previous studies in cancer as well as population genetics have reported great successes in identifying disease susceptibility genes via germline SNVs [4-11] and common SNPs [12, 13]. Yet with the accumulation of knowledge on the impacts of mutations on disease predisposition, disease etiology[14] as well as the development of new clinical applications, there is a clear demand for identifying and analyzing low frequency SNVs in basic biological research and clinical applications [15-19].

1.1.1 Low Frequency SNVs in Cancer Research and Clinical Applications

In cancer research, a widely accepted notion is that cancer is a complex disease arising from sequentially accumulation of somatic mutations, which leads to the transformation of normal cells to cancer cells [20, 21]. Thus somatic mutations are the key for us to understand carcinogenesis and also seek proper treatments. However, identifying somatic mutations from tumor samples is more challenging than germline mutation detection using purified peripheral blood. The first reason is the lower frequencies of these somatic mutations since tumor samples from biopsies are often of low purity. The low purity is the result of normal cells contamination [22, 23], as well as the highly heterogeneous nature of the tumor cells, which are a mixture of multiple genetically different tumor subpopulations [24, 25]. Previous work summarized the estimated tumor purity from existing major cancer studies, in which lung cancer has a large number of samples with purity between 20% and 40% and the purity for some

samples are even less than 20% [22]. Different levels of heterogeneity are observed in tumor. The Intratumor and the intercellular genetic heterogeneity are the main sources of complication for identification of low frequency mutations. In addition, the tumor samples may have lower quality. An increased background mutation rate is expected for cancer biopsy specimens that are formalin-fixed and paraffin-embedded (FFPE) due to cross-linking [26, 27]. Thus, efficiently distinguishing low frequency somatic mutations from background errors carries great significance in cancer research. It is also important in clinical applications, since it enables the early diagnosis, cancer progression monitoring and relapse identification, which are essential components of cancer treatment.

Besides, low frequency mutation detection is also desirable in newly developed applications. The recent discovery of circulating tumor DNA (ctDNA) gained much attention from cancer researchers and clinicians since contrast to traditional tumor biopsy, which is invasive and can only offer a snapshot of the tumor genetics landscape at certain checkpoints, ctDNA based 'liquid biopsy' [28] is non-invasive and can be done repeatedly for close monitoring of early sign of relapse or metastasis [29-32]. However, ctDNA only represents a small percentage of all blood sample DNA [33]. A previous research [34] reported for some advanced cancers, ctDNA is about 1~10% of blood DNA.

1.1.2 Low Frequency SNVs in Population-Genetics Research

Single nucleotide polymorphism (SNP) is an SNV occurring within at least 1% a population [35]. Information about polymorphic positions in the genome and the analyses on frequencies of variant alleles in various populations are the key inquiries of population genetics. To increase the power of population genetic studies, estimating allele frequencies from a large number of population samples is desirable due to higher accuracy. Individually sequencing a large number of samples is usually cost-prohibitive,

thus sequencing pooled DNA samples [36-38] as a cost-effective alternative was developed and also proved to generate more accurate allele frequency estimations [39-41] than individual sequencing at similar cost. However, pooling larger number of individuals also brings the challenge of distinguishing sequencing errors from low frequency alleles. Further, with the paradigm shift in complex diseases studies from 'common disease-common variants' (CDCV) to 'common disease-rare variants' (CDRV) [42-44], the importance of reliably identifying low frequency (0.5~5% minor allele frequency or MAF) to rare (MAF < 0.5%) variants is again pinpointed.

1.2 NGS Protocols, Applications and Limitations

Determining the sequence composition is a fundamental task in biomedical researches. To determine the sequence of base pairs that make up human DNA, the Human Genome Project was launched in 1990 and took $3 billion and 13 years to complete. The monumental project was accomplished with Sanger sequencing, which is considered the first generation sequencing technology. Despite many technical improvements made to the technology, after dominating the sequencing industry for more than two decades, Sanger sequencing could not keep pace with the great demand for cheaper, faster and more accurate sequencing of a large number of human genomes. This demand catalyzed the development of next-generation sequencing (NGS), which performs massively parallel sequencing and allows the entire human

genome to be sequenced within days. The much-improved sequencing technology revolutionized genomics and genetics researches and applications. To fully utilize its power, a clear understanding of its design and basic principles, as well as its power and limitations is indispensible. This is the key for low frequency SNV detection, since the task requires sensitively and specifically distinguishing true SNVs with close to sequencing error rate frequency from sequencing artifacts. In this section, the NGS DNA sequencing protocols are first introduced, then sequencing read alignment methods and complications are briefly introduced. The error sources of NGS platforms are summarized and the efforts trying to mitigate the problem are also introduced.

1.2.1 NGS DNA Sequencing Experimental Protocols

Since the release of the first NGS sequencer by 454 Life Science, during the past decade, many NGS platforms based on different technical protocols have been developed and released. Among those platforms, the leading benchtop sequencers for targeted gene panel or small genome sequencing are Illumina MiSeq, Ion PGM and Ion Proton. While the leading population- and production-scale sequencers designed for large number of whole genome, exome or transcriptome sequencing are Illumina HiSeq series.

Taking Illumina sequencer as an example, a DNA sequencing experiment includes the following steps [45]:

1. Library Preparation – for whole genome sequencing, the genomic DNA sample is randomly fragmented by sonication or nebulization, followed by 5’ and 3’ adapter ligation. Adapter-ligated fragments are PCR amplified and gel purified. For amplicon-based targeted sequencing, custom amplicon probes hybridize to flaking regions of interest in unfragmented genomic DNA to capture the desired sequences. Then PCR adds sequencing adapters to the amplicons and the amplicon library is ready for amplification.
2. Library Amplification – The sequencing library is loaded in to a flowcell. Adapter-ligated DNA fragments are separated into single strands. The surface of the flowcell is bounded by millions of oligos

complimentary to the library adapters, thus the single stranded library fragments are captured. Then each captured fragment is amplified into a clonal cluster via bridge amplification.

3. Sequencing – Illumina uses sequencing-by-synthesis (SBS) technology. At each cycle, 4 types of fluorescently labeled nucleotides are added, the ones complementary to the template DNA are added and the emission from each cluster on the flowcell is recognized and recorded by the optical imaging system. The bases incorporated and the qualities are determined from the emission wavelength and intensity data.
4. Data Analysis – Sequencing reads are aligned to a reference genome. Different variant calling algorithms can be applied on the aligned sequencing data.

For Ion Torrent sequencers, a DNA sequencing experiment shares the above 4 general steps but the technical details are different [46]. For library amplification, emulsion PCR is used. Adapter-ligated DNA fragments are separated into single strands and then are captured by beads under conditions favoring one template per bead. The DNA-bead complexes are then mixed with oil-aqueous emulsion to create individual droplets that encapsulate these DNA-bead complexes. These droplets are also called microreactors in which PCR amplification is performed. In the sequencing step, millions of beads flow across the Ion semiconductor chip, each depositing into a well. Then the chip is flooded with a sequence of the 4 nucleotides. Whenever a nucleotide incorporates a single stranded DNA, a hydrogen ion is released and changes the pH of the solution in the well. An ion sensitive layer beneath the well detects the change in pH and converts that to voltage and thus the base is detected. This technology is called Ion semiconductor sequencing.

1.2.2 NGS Read Alignment

NGS platforms generate large number of short reads. The majority of high-throughput NGS platforms now can generate single or paired-end reads with 100 to 300 in length, with Ion PGM system capable of generating millions of 400-base length reads. The relatively lower-throughput sequencer from 454 can generate reads up to 1000

nucleotides in length. In terms of throughput, Ion Torrent benchtop sequencer Ion PGM 318 chip can generate 4-5.5 million reads while Ion Proton can generate 60-80 million reads. Illumina benchtop sequencer MiSeq can generate 25 million reads per run while population- and production-scale sequencer HiSeq series can generate 6 billion reads per run.

Fast and accurately aligning enormous amount of short reads back to the reference genome is the key issue for NGS data analysis. To solve this problem, different strategies had been tested. From hashing the short reads [47-49], hashing the reference genome [50, 51] to Burrows-Wheeler transform (BWT) used by string matching theory, the speed, memory usage, error tolerance had been greatly improved. For Illumina sequencing data, BWT based Burrows-Wheeler aligner (BWA) [52] is the most widely used tool. For Ion Torrent data, TMAP from Torrent Suite Software is used since it optimized the modules and parameters to adapt to flexible read length. One major difficulty of read alignment is aligned reads from low-complexity reference genome regions. In addition, sequencing errors, including mismatches and micro-insertions and deletions (INDELs) can also complicate the alignment.

1.2.3 NGS Error Profiles and Error Reduction Methods

With the ever-increasing importance of NGS in genomics and genetics, the significance of accounting for experimental errors is also more prominent, especially in the application of identifying low frequency variants where the variant allele frequency could be close to or below sequencing error rate (0.1~1% for most platforms [53]).

Starting from sample preparation, all steps can potentially generate errors [54]. In sample preparation step, nucleic acid degradation and FFPE crosslinking induced errors [55], as well as alien sequence contamination are the main sources. In library amplification step, PCR amplification errors [56] may be recognized as SNVs in the

subsequent SNV calling. During the sequencing step, all platforms show error profiles related to GC and/or AT content as well as homopolymers [57], and elevated error rates in low-complexity regions. In terms of the relationship between GC content and error rate of different platforms, Illumina and Complete Genomics platforms are more sensitive to GC content differences. For homopolymer length, except for Pacific Biosciences considered as 'third generation sequencing' technology [58], all the other platforms show elevated error rates when homopolymer length is larger than 10. Nevertheless, different platforms are more vulnerable to different types of errors due to the differences in the underlying technologies. Illumina and Ion Torrent sequencers are based on totally different sequencing biochemistries. For Illumina sequencers, substitution error is the major error source [59] and Nakamura et al. Identified sequence contexts that tend to trigger these errors [60]. Ion Torrent tends to have more indel errors around homopolymer sequences [61] and thus tends to generate false SNV calls due to erroneous alignment.

To mitigate the NGS errors, many researchers have reported successes in reducing NGS errors by improved experimental protocols, especially the library preparation and amplification steps. "Barcoding" strategy has been discussed in several papers where a mutation is confirmed only if it appears in multiple read groups distinguishable by the barcodes [16, 62]. Circle sequencing improved the idea of independent mutation confirmation from multiple read groups by removing the need of adding barcodes [63]. Duplex sequencing approached the error reduction by requiring strand concordance on the mutations [18, 64]. However, most existing NGS data are generated from standard experimental protocols without the specially designed steps implemented in barcode, circle and duplex sequencing protocols described above, thus how to effectively distinguish sequencing artifacts and errors from low-frequency SNVs is an important topic in bioinformatics. In terms of bioinformatics approaches, many

researchers use filters, including requirements on sequencing read depth, base quality, mapping quality, strand bias, variant quality and mutation density [65]. Also, some researchers proposed using replicates [54] to reduce errors. SNVs are called from all replicates and then classified. SNVs agree among all the replicates are treated as concordant, and discordant if not. Concordant SNVs are more likely to be true SNVs rather than sequencing errors. The replicates could be technical, biological and cross-platform. Taken biological replicates as an example, by plotting fraction of concordant and discordant SNVs on different thresholds of different filter metrics, such receiver–operator characteristic curves can help evaluate the efficiencies of different filters and determine the threshold as well.

1.3 NGS Based Low Frequency SNV Detection: Challenges and Existing Efforts

The main challenges are how to reliably measure low frequency SNVs and how to distinguish from sequencing artifacts. Targeted deep sequencing can generate NGS data with per base coverage up to thousands or even higher, thus are more likely to capture low frequency SNVs. Amplicon based PCR target capture assay is the most common choice, and the target region usually ranges from tens (several kilobases) to hundreds of genes (up to several million bases). The development of benchtop sequencing systems such as Ion Torrent PGM, Ion Proton and Illumina MiSeq and NextSeq series, greatly promotes targeted sequencing and the development of target panels. However, tumors are genetically heterogeneous and often contain normal/stromal cells, which render some low-abundance somatic mutations close to or even below NGS detection limit. Targeted sequencing generates deep coverage on those loci but the amplification step keeps the original mutant to wildtype ratio in the tumor samples, with potential bias toward the wildtype allele. Methods that can selectively amplify the mutant allele have the potential to enrich these subtle signals.

COLD-PCR (co-amplification at lower denaturation temperature-PCR) [66-69] is a modified PCR protocol that identifies and enriches low-level mutant alleles in the presence of excess wildtype alleles, thus enabling the downstream analysis to identify real low-frequency variants. However, the feasibility to generalize COLD-PCR to a large number of sequences and sequences of various nucleotide compositions need to be carefully evaluated. Moreover, there is a recent report [70] of a sophisticated experimental protocol using target enrichment by sequential rounds of hybridization with biotinylated oligonucleotides, together with duplex sequencing described in section 1.2.3 to increase the accuracy of calling rare variants. However, it is primarily applicable to small genomic intervals of the size of a single gene. Thus despite the promising results demonstrated from current protocol developments, more efforts in generalizing and standardizing these modified protocols are required for broader applications.

In terms of bioinformatics methodologies, existing tools, such as VarScan2 [71], Strelka [72], and mutect [73], are mainly designed to target variants with lowest allele frequencies at 5% level for a whole exome or several hundreds of targeted genes sequenced with average depth around hundreds. Several studies focus on a small number of hotspot cancer genes with ultra-deep sequencing (greater than 10,000x in depth) [74, 75] for pushing down the detection limit. However, such methods usually take ad hoc filtering approach, and are designed to target variant identification within a small genomic region, usually less than 20,000 nt. In addition, existing methods typically use base quality to derive the base call error rate for each location assuming equal substitution error rates, and/or using an empirical mutational rate to derive the posterior probability or likelihood ratio of a location being a somatic mutation rather than a germline variant. Since the common parameters used didn’t consider differential error profile at different genomic loci across the targeted regions, such method is suboptimal in sensitively detecting variant with allele frequency close to intrinsic sequencing error

rate. Thus, additional ad hoc trimming, filtering and thresholding are often required to remove the variants with lower qualities. Such strategy significantly limits the generalizability of the analysis methods. Therefore, to our knowledge, no existing methods can reliably detect SNVs at close to 1% allele frequency using data from standard sequencing protocols targeting hundreds of genes.

1.4 Appropriate Benchmarking for Low Frequency SNV Detection Methods

Defining a gold standard benchmark for SNV detection is challenging since the benchmark needs to be completely characterized to include all real SNVs as well as to exclude false positive calls. Such a task is even more challenging for low-frequency somatic SNV detection since no tumor genome has been completely characterized and the depths of existing data usually are not deep enough to enable identifying low-frequency SNVs. To address these problems, there are in general 4 types of approaches to simulate cancer genomes for benchmarking: (1) de novo simulation reads and mutations based on previously learnt sequencing error profiles on the basis a reference genome [76-80], (2) admixture of existing sequencing data from multiple samples at various percentages [23, 81], (3) bridging (1) and (2) where cancer genome reads are derived by modifying pre-existing alignments at desired frequencies and realigning the modified reads [82] and (4) mixing DNA samples with known genotypes at designed percentages and then sequence the DNA mixture [83]. These methods all have their own merits and demerits. Method (1) is cost-efficient since once the error profiles are learnt, it can generate new simulated sequencing data without actual sequencing cost. And it is flexible since simulated data can be generated for different platforms based on different error profiles. However the major problem with this method is that it cannot recapitulate biases and error profiles if they had not been well defined and characterized. This is a serious problem that limits its application value since different

combinations of sample preparation and sequencing technology may demonstrate differential biases and error profiles. Method (2) successfully avoids the problem in (1) by operating on existing sequencing data, thus the sequencing biases and error profiles are well preserved. However, there are concerns about method (2) arguing it is biased toward SNVs already detectable [82]. In addition, the allele-frequencies derived from in-silico mixing step may not be an ideal representation of the actual biological cell subpopulations. The reason is the mechanisms causing such variations may be far more complicated than the in-silico mixing schemas. Method (3) tried to combine (1) and (2) to take advantages of both, however, this method ignores the context specific nature of sequencing error profiles [57, 60, 84]. Further, the rationale of retaining the same base qualities after changing bases at the same locus is open to doubt. Method (4) preserves the original sequencing biases and error profiles, especially the sequence context related errors. It can also provide information to evaluate the agreement between the observed allele frequencies and the actual abundance of corresponding cell populations, which is useful for determining the variability of allele frequencies estimated from sequencing data. In addition, compared with in-silico mixing, the DNA-mixing-followed-by-sequencing approach allows characterization of potential bias toward wildtype allele, which may affect allele frequency estimation. For the purpose of benchmarking low-frequency SNV callers, we choose method (4).

1.5 Objectives

The main objective of this thesis is to develop a computational framework to reliably identify low-frequency SNVs for applications in cancer or population genomics and genetics study. To serve the main objective of computational methodology development, three sub objectives were derived: design suitable benchmark for model training and testing, model sequencing errors to establish the lowest detection boundary

and decide the optimal classification boundary between true SNVs and technical artifacts to refine SNV calls.

Chapter 2 describes in detail a novel experimental and computational modeling framework, which designed specific modules for each of the aforementioned sub objectives. The framework named RareVar aims to push the detection limit to allele frequencies as low as 0.5-1% under standard sequencing experiment protocols. This would significantly improve the sensitivity with which rare somatic mutations can be detected. The experimental part includes a strategy to construct benchmark tumor DNA samples containing thousands of SNVs with a wide range of allele frequencies yet enriched with low frequency variants (0.5%-3%). The benchmark tumor DNA samples are amplified and sequenced using the same protocol as the primary tumor samples, and are further used to construct a statistical model for deriving the background sequencing error rates that are specific to different genomic loci. This benchmark sample is further used for constructing a machine-learning-based model for variant recalibration. We evaluated the performance of RareVar together with several existing tools on an independent test benchmark. This analysis showed RareVar is more sensitive than other tools for variants at 3% or less allele frequencies.

Chapter 3 described the ongoing project of applying our low-frequency SNV calling framework RareVar on studying mutational drift/enrichment of reprogrammed breast tumor cells.

Chapter 4 explored the potential to improve the performances on identifying candidate SNVs with close to sequencing error rate frequencies by implementing more sophisticated statistical models for sequencing error characterization.

Chapter 5 evaluated the generalizability and adaptiveness of the position specific sequencing error model. Instead of Ion Proton sequencing data used in previous

chapters, the model was tested on Illumina MiSeq platform, which utilized completely different biochemistries.

Chapter 2. RareVar: a Framework for Detecting Low Frequency Single Nucleotide Variants

2.1 Overview of RareVar Framework

The RareVar protocol includes five major components: benchmark sample design, target region amplification and sequencing, position specific sequencing error modeling (PSEM), variant identification, and machine-learning-based variant recalibration (Figure 1). A training benchmark sample was designed to contain a set of mutations at known allele frequencies in the desired capture regions. This benchmark sample was sequenced in parallel with the samples of interest using the exact same capturing and sequencing protocol and thus serves as a calibration set to evaluate the accuracy of the sequencing and analysis pipeline. The non-SNV loci in the benchmark sample provide data for PSEM on genomic features that distinguish low frequency SNVs from sequencing errors, while the known SNVs allow further adjustment of machine-learning algorithms to recalibrate the variant calls based on features from the particular experimental procedures.

In this chapter, the usage of RareVar framework is demonstrated under the scenario of detecting somatic SNVs from paired normal-tumor samples. Thus, both training and testing benchmarks contain 2 samples, one mimicking the normal sample and the other one mimicking the tumor sample. In the training benchmark, instead of using invariant loci data from tumor sample as the training data for position specific error modeling, the invariant loci data from normal sample sequencing data are used. The reason is we want to take advantage of the paired normal-tumor design, since the normal sample is less likely to contain potentially missed SNVs compared with the tumor sample that is generated by mixing DNAs from multiple individuals as described in section 2.2.1. The SNV loci sequencing data in the tumor sample constitute the training data for machine-learning-based variant recalibration.

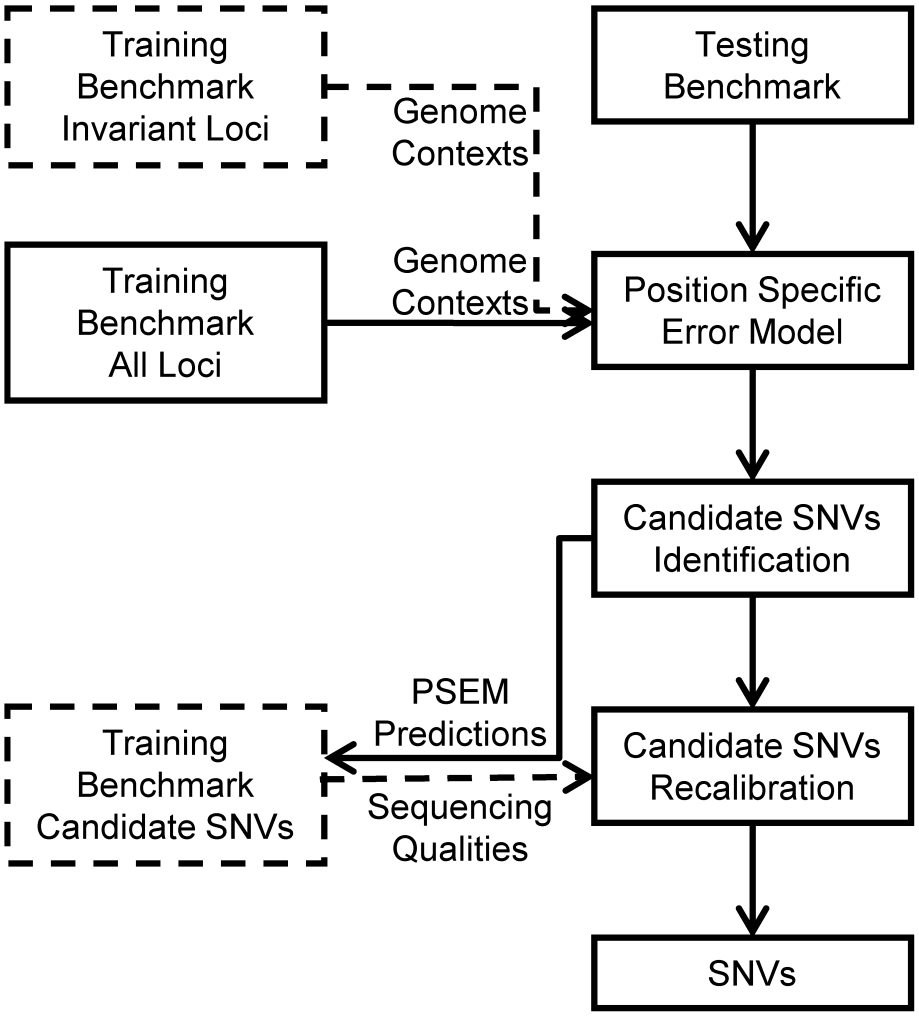

Figure 1 RareVar framework overview. During the training phase, genome contexts of invariant loci are used to train a position specific error model (PSEM). Then the genome contexts of all loci are fed to PSEM and the resultant predictions comprise the candidate SNV loci. Sequencing qualities of those candidates are used to further calibrate their fidelity. Actual data involved in model training are highlighted in dashed lines and boxes. During the testing phase, testing benchmark data go through the trained PSEM and recalibration model to generate high confidence SNVs.

2.2 Materials and Methods

2.2.1 Benchmark Design

A total of 22 DNA samples from the 1000 Genomes Project were selected. The genotype information is available for the selected individuals [85]. Two sets of 18 samples were used, one for the training benchmark set and the other for the testing benchmark set (Table 1). For paired normal-tumor design, one sample was chosen as the normal sample, and then the 18 samples were pooled together at different percentages to mimic the tumor sample. Details are described below.

The goal for the training benchmark tumor sample mixing design was to maximize the number of low frequency (0.5-3%) SNVs in the target regions. To achieve this goal, two steps were utilized. First, among one set of 18 DNA samples, we identified the one that has the largest number of overlapping SNVs with other samples in the target regions, which is NA11993 shown in Table 1. The SNVs from this sample were used to represent the germline mutations from normal/stromal cell population for somatic mutation identification. These SNVs are referred as “germline” SNVs and the sample is referred as “normal” sample in the later description. Second, for the tumor sample, the other 17 samples were mixed at varying concentrations (1% to 10%); samples with larger number of unique SNVs in the target regions were assigned lower concentrations (Table 1). The previously chosen normal sample represents normal cell population in the pooled tumor sample. Similarly, the testing benchmark tumor sample was designed by mixing another set of 18 DNA samples at concentrations different from the training samples. Four of the 18 samples in the testing benchmark were not in the training benchmark (Table 1 last 4 rows), and the three samples with higher number of unique SNVs were assigned 1% mixing percent while the remaining sample NA12878 was assigned the highest mixing percent. The SNVs from the new samples in the testing benchmark tumor sample comprised the subset to monitor potential model over-fitting

from both the position specific error model and the machine-learning step, where SNVs from the samples used in both training and testing comprised the subset to evaluate the performances on independent sequencing runs. The DNA mixing strategy for both training and test benchmark tumor samples are shown in Table 1.

2.2.2 Sequencing

The targeted regions for benchmark datasets included all exons of 409 known cancer-related genes, totaling about 1.7 million bases. For library construction, targeted sequences were captured by ~ 16,000 amplicon primer pairs from Ion AmpliSeq Comprehensive Cancer Panel. The average length of amplicons is 155bp. The library was prepared using Ion AmpliSeq Library Kit 2.0. Sequencing was carried out on Ion Proton system, and the data was aligned to Human Genome reference hg19 by TMAP in Torrent Suite Software version 4.4.2. On target, uniquely mapped and mapping quality >= 40 reads were used in the following analyses.

Table 1 Design of tumor samples for training and testing benchmarks. Individual NA11993, shaded, was used as the normal sample in the training set; while shaded individual NA12878 was used as the normal sample in the testing set. 'NA' means not used.

	Training Set Tumor Sample		Testing Set Tumor Sample	
ID	**Number of Unique SNPs**	**Mixing Percent**	**Number of Unique SNPs**	**Mixing Percent**
NA11993	37	0.47	NA	NA
NA18507	86	0.10	70	0.01
NA12155	33	0.08	NA	NA
NA18563	37	0.06	NA	NA
NA12144	32	0.03	NA	NA
NA12750	42	0.03	43	0.10
NA12751	42	0.03	38	0.06
NA07000	39	0.03	43	0.06
NA18987	44	0.03	44	0.03
NA18965	48	0.02	43	0.08
NA12872	51	0.02	45	0.03
NA18622	44	0.02	44	0.03
NA18853	72	0.02	60	0.01
NA18526	48	0.02	58	0.01
NA18870	109	0.01	102	0.02
NA18502	107	0.01	90	0.02
NA18871	99	0.01	85	0.02
NA18501	95	0.01	75	0.02
NA19239	NA	NA	86	0.01
NA12878	NA	NA	41	0.46
NA19238	NA	NA	71	0.01
NA19092	NA	NA	65	0.01

2.2.3 Position Specific Error Rate Modeling

In order to evaluate how genome contexts at specific genomic loci affect sequencing accuracy, a Poisson distribution generalized linear model (PD-GLM) was applied to model the relationship between specific genome contexts and error rates as shown in Equation (1),

$$\log\left(E\left(n_{l,b,s}\right)\right) = \log\left(d_{l,s}\right) + \alpha + \overrightarrow{\beta'} * \overrightarrow{X_{l,b,s}} \quad (1)$$

where $n_{l,b,s}$ is the number of reads in location l within the target regions that support non-reference base b on strand s (forward or reverse), $d_{l,s}$ is the depth of sequencing at location l on strand s, and $\overrightarrow{X_{l,b,s}}$ is a vector of co-variants that describes different aspects of genomic context surrounding the candidate loci. In addition, α and $\vec{\beta}$ are the intercept and coefficients of the regression for the co-variates, which indicate the contribution of each factor to the sequencing error rate. The training data for PD-GLM contain the loci that have a depth within 25% to 75% quantile and alternative allele frequency no more than 1.5%, totaling ~ 5 million records.

PD-GLM integrated 9 genome context features (Table 2) previously reported to be related to sequencing errors [57, 61], including alternative base substitution types, the nucleotides immediate upstream and downstream of the variant loci, the percentage of GC nucleotides in the nearby region by extending 50 bases upstream and 50 bases downstream of the target nucleotide. In addition, features related to homopolymer of the loci were also considered, including the length of the closest homopolymer, the distance from the SNV to the closest homopolymer (defined as the number of bases from the target nucleotide to the closest base in the homopolymer, specifically, the homopolymer could be upstream/downstream or could contain the locus of interest) and the fraction of

bases within homopolymers in the nearby region by extending 15 bases upstream and 15 bases downstream of the target nucleotide. The homopolymer features are designed to capture the intuition that other than nucleotide contexts (substitution, upstream, downstream bases and GC content), sequencing data for a locus tend to be erroneous if it is near the boundary of one or more long homopolymer(s).

Table 2 Definition of features in the PSEM step and summary of regression.

Features	Definition	Degrees of Freedom[1]	Covariates[2]	Estimated Coefficients	Standard Error[3]	P Value[4]
NA	Intercept only model, containing the baseline for each feature. Specifically: substitution is A > C, upstream and downstream bases are both A; GC, hmer_dist, hmer_len, hden, hrun_op and alt_up_down_eq are all 0.	1	Intercept	-11.030	0.0079	< 2e-16
substitution	Change from reference base to alternative base; there are 12 possible values. A > G means reference base A to alternative base G.	11	A > G	1.621	0.0049	< 2e-16
			A > T	0.123	0.0063	< 2e-16
			C > A	0.046	0.0065	1.8e-12
			C > G	-0.184	0.0070	< 2e-16
			C > T	1.399	0.0051	< 2e-16
			G > A	1.326	0.0051	< 2e-16
			G > C	-0.071	0.0068	< 2e-16
			G > T	-0.022	0.0066	9.1e-04
			T > A	0.190	0.0062	< 2e-16
			T > C	1.633	0.0049	< 2e-16
			T > G	0.134	0.0061	< 2e-16
upstream base	Immediate upstream base (4 possible values: A,C,G,T).	3	C	0.115	0.0026	< 2e-16
			G	0.247	0.0025	< 2e-16
			T	-0.127	0.0027	< 2e-16

Features	Definition	Degrees of Freedom[1]	Covariates[2]	Estimated Coefficients	Standard Error[3]	P Value[4]
downstream base	Immediate downstream base (4 possible values: A,C,G,T).	3	C	0.475	0.0027	< 2e-16
			G	0.308	0.0027	< 2e-16
			T	0.118	0.0028	< 2e-16
GC content	Percent of GC bases within a 101 base window that extends 50 nucleotides both upstream and downstream.	1	GC	0.005	0.0001	< 2e-16
distance to the closest homopolymer [5] base	Number of nucleotides to the closest base of the homopolymer within a window that extends 15 bases both upstream and downstream (possible values 0 to 13, 15 for no homopolymer within the window).	1	hmer_dist	-0.009	0.0003	< 2e-16
length of the closest homopolymer	Length of the closest homopolymer within a window that extends 15 bases both upstream and downstream (possible values 0, 3, 4 to 31).	1	hmer_len	0.081	0.0009	< 2e-16
homopolymer bases percentage	Fraction of bases within a 31 bases window that are in homopolymers (possible values 0, 3/31, 4/31 to 1). The window extends 15 bases both upstream and downstream.	1	hmer_percent	0.193	0.0095	< 2e-16
overlap with homopolymer	Whether the locus of interest is within a homopolymer, 1 means yes, 0 means no.	1	hmer_op	0.375	0.0026	< 2e-16
upstream or downstream base shift	Whether the alternative base is the same as the immediate upstream or downstream base, 1 means yes, 0 means no.	1	alt_up_down	0.690	0.0019	< 2e-16

1: Degrees of Freedom: The number of covariates for each feature in the PD-GLM model. For categorical features, this is the number of possible levels minus 1 while for numerical features degrees of freedom equal 1.
2: Covariates: symbols for all features used in PD-GLM. These are the variable names used in the generalized linear model.
3: Standard Error: the standard error of the estimated PD-GLM coefficients.
4: P Value: significance of each covariate.
5: homopolymer: a consecutive sequence of at least 3 identical bases.

2.2.4 Variant Identification

We applied a Bayesian-based approach for identifying variants with low allele frequencies, based on the number of reads supporting alternative allele at each specific genomic locus, and its estimated position-specific error rate. For each candidate variant locus, a Bayes factor was calculated by comparing the likelihood ratio of two competing models - M_E and M_V. M_E represents the model that the number of alternative reads follows 'sequencing error distribution' - PD-GLM estimated position-specific sequencing error. Whereas M_V represents the model that the number of alternative reads follows the 'targeted lowest identifiable frequency distribution' - a SNV at the frequency of the targeted lowest identifiable allele frequency more than PD_GLM predicted error. In this study, our targeted lowest frequency is f = 0.5%. In Equation (2) $n_{l,b,s}$ and $d_{l,s}$ remain the same as in Equation (1). In addition, $\lambda_{E,l,b,s}$ and $\lambda_{V,l,b,s}$ represent the expected number of alternative reads assuming the candidate locus is not an SNV ($n_{l,b,s} \sim Poi(\lambda_{E,l,b,s})$), and is an SNV with the lowest intended identifiable allele frequency ($n_{l,b,s} \sim Poi(\lambda_{V,l,b,s})$),), respectively. An observed substitution type in a location is considered a SNV candidate if Bayes factors $BF_{l,b,s}$ for both strands are greater than 100. This threshold implies that on each strand it is 100 times more likely that a specific position is a variant than that it is a sequencing error. We evaluated the precision, recall and F1 scores at different Bayes factor thresholds, including any number that is a power of 2 within 2 to 512, together with 10, 50, 100, 200, 300, 400 and 500. Figure 2 upper panel showed recall dropped with increasing thresholds while the precision increased. The lower panel showed harmonic mean of precision and recall – F1 score increased with bigger thresholds. However, the increase in F1 score began to significantly slow down around 64 to 128, as highlighted by the tangential line at 100. Thus the performance gain by setting more stringent threshold became smaller. Since the variant identification step aims at identifying SNV

candidates, we chose 100 as the threshold to efficiently gain increase in precision and left the candidate refinement to the next step.

$$BF_{l,b,s} = \frac{Pr\left(n_{l,b,s}\middle|M_V\right)}{Pr\left(n_{l,b,s}\middle|M_E\right)} = \frac{\frac{\sum_{k=0}^{n_{l,b,s}} \lambda_{V,l,b,s}^{k} e^{-\lambda_{V,l,b,s}}}{k!}}{1 - \frac{\sum_{k=0}^{n_{l,b,s}} \lambda_{E,l,b,s}^{k} e^{-\lambda_{E,l,b,s}}}{k!}},$$

$$\lambda_{V,l,b,s} = \lambda_{E,l,b,s} + d_{l,s} * f \quad (2)$$

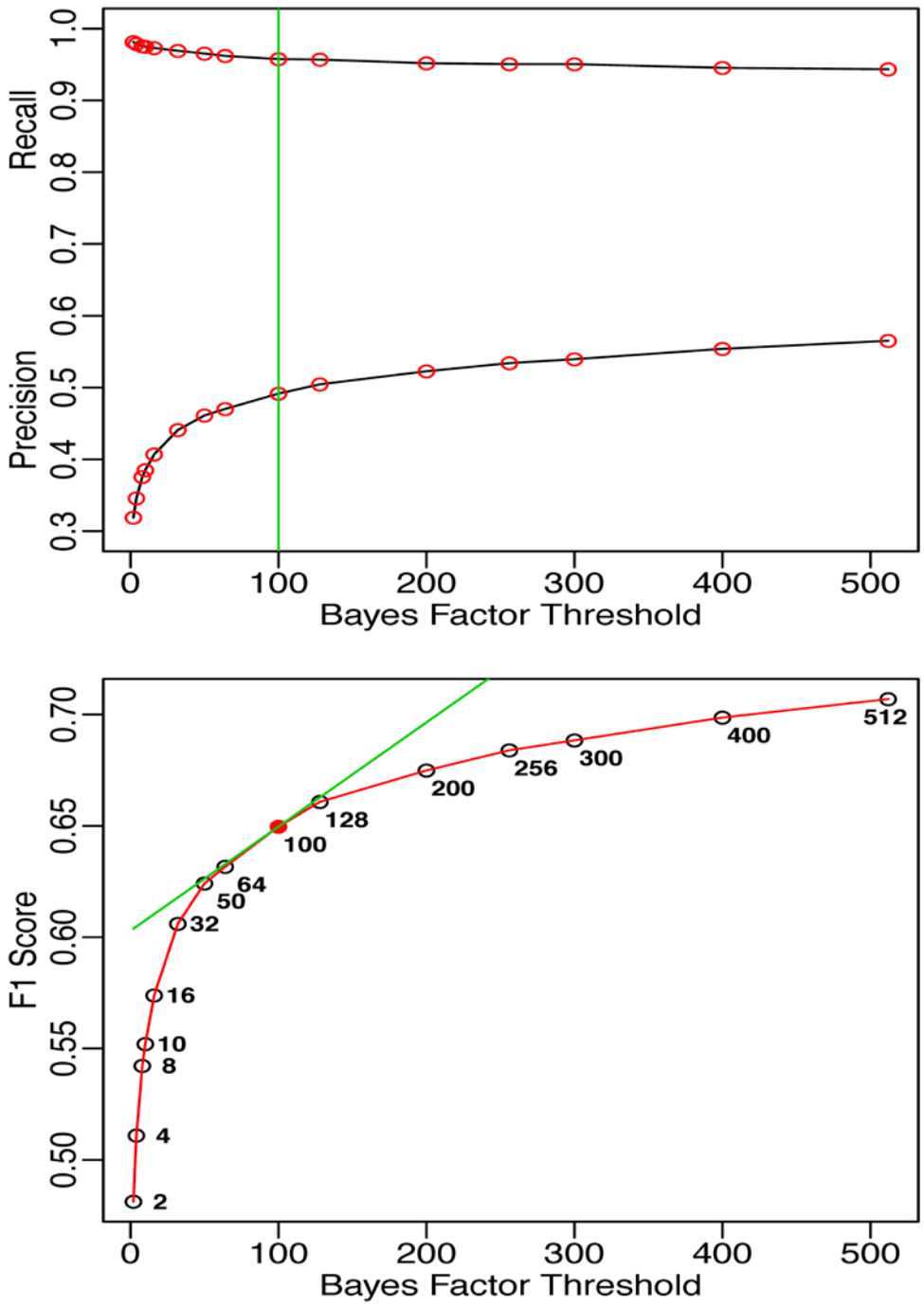

Figure 2 Performance metrics at various thresholds of Bayes factor.Upper panel shows the precision (bottom line) and recall (top line) at various thresholds. Lower panel shows the F1 score at various thresholds. The green line in the lower panel is the tangent line of the curve at threshold 100.

2.2.5 Machine-Learning Based SNV Calibration

SNV candidates at the lower frequencies from the variant identification step still contain a large number of false positives. It has been observed previously that sequencing-related measurements, such as sequencing and alignment quality, have a strong influence on the accuracy of variant identity [65]. Instead of setting up a series of hard filtering criteria for different measurements, a strategy utilized by many earlier methods, we adopted a machine learning-based approach to derive an optimal classification boundary between false positives and true ones by simultaneously modeling multiple measurements. This strategy takes the advantage of the benchmark dataset we constructed, since true and false positives of the identified candidate variants in these samples are known. Since the PSEM model focused on the genomic features surrounding the variant loci, in this second step, we further explored the measurements related to the experimental and analytical steps. Many of these features have been reported useful in ruling out false positives in previous studies [65, 71, 72, 86] (Table 3). The features included in the machine-learning model can be grouped into the following generic types: sequencing, alignment, amplicon structure and genome context related features from PSEM. Information gain (IG) [87] was used to rank the classification power of all features and is defined in Equation (3), where C is the two classes: true SNVs or noise; $H(C)$ is the information entropy for all classes, $H(C|feature)$ is the conditional information entropy for all classes given a feature; x is a categorical feature with k levels, where numerical features are discretized with Fayyad & Irani's MDL method [88].

Table 3 RareVar: features considered in variants recalibration step. Features are ranked based on column 'Information Gain'. Genome context features are from the PSEM step.

Feature	Explanation	Category	Information Gain	Source
Allele_Freq	Allele frequency	Sequencing.	0.363	RareVar
MisMatch_Percent	For reads containing alternative base, the percent of reads also containing other mismatches within a 11 base window.	Alignment.	0.300	RareVar
MAPQ_Alt_RefDiff	Difference of average mapping quality between reads containing alternative and reference.	Alignment.	0.220	RareVar
MAPQ_Avg_Alt	Average mapping quality for reads containing alternative base.	Alignment.	0.197	RareVar
BaseQ_Avg_Alt	Average base quality for alternative bases.	Sequencing.	0.190	RareVar
MAPQ_RankSum	Rank sum test of mapping quality.	Alignment.	0.177	GATK
BaseQ_RankSum	Rank sum test of base quality.	Sequencing.	0.130	GATK
BaseQ_Alt_RefDiff	Difference of average base qualities between alternative and reference bases	Sequencing.	0.117	RareVar
Fisher_SB	Fisher exact test of strand bias.	Sequencing.	0.106	GATK
AbsDiff_StrandAF_Percent	Absolute value of the difference in allele frequencies between two strands	Sequencing.	0.102	RareVar
Substitution	Change from reference base to alternative base.	Genome context.	0.073	RareVar
Strand_OR	Odds ratio of reads supporting alternative and reference alleles in two strands.	Sequencing.	0.068	RareVar
Hmer_dist	Distance to the closest homopolymer base.	Genome context.	0.064	GATK
Alt_up_down	Upstream or downstream base shift	Genome context.	0.053	RareVar
BaseQ_Avg_Ref	Average reference base quality.	Sequencing.	0.049	RareVar
Hmer_op	Overlap with homopolymer.	Genome context.	0.047	RareVar
Feature	**Explanation**	**Category**	**Information**	**Source**

			Gain	
Hmer_len	Length of the closest homopolymer.	Genome context.	0.046	RareVar
Fwd_WDis	Forward strand weighted distance to amplicon ends.	Amplicon.	0.035	RareVar
Fwd_Read_Percent	Percent of forward strand reads supporting alternative allele.	Sequencing.	0.034	RareVar
Rev_WDis	Reverse strand weighted distance to amplicon ends.	Amplicon.	0.032	RareVar
ReadPos_RankSum	Rank sum test of alternative allele position in reads.	Sequencing.	0.032	GATK
Up_base	Immediate upstream base.	Genome context.	0.030	RareVar
Down_base	Immediate downstream base.	Genome context.	0.029	RareVar
HDen	Homopolymer density, the percent of homopolymer bases within a 31 bp window.	Genome context.	0.025	RareVar
MAPQ_Avg_Ref	Average mapping quality for reads containing reference allele.	Alignment.	0.016	RareVar
Quality_Depth	SNP confidence normalized by unfiltered depth of snp samples.	Sequencing.	0.015	GATK
Avg_NMM_perRead	Average number of other mismatches in reads containing alternative allele within a 11 bases window.	Alignment.	0.014	RareVar
RMS_MAPQ	Root Mean Square of the mapping quality of reads.	Alignment.	0.01	GATK
GC Content	Percent of GC bases within a 101 base window that extends 50 nucleotides both upstream and downstream.	Genome context.	0.000	GATK

$$IG = H(C) - H(C|feature)$$

$$= -\sum_{j=1}^{2} P(c_j) log P(c_j) + \sum_{i=1}^{k} P(x_i) \sum_{j=1}^{2} P(c_j|x_i) log P(c_j|x_i) \quad (3)$$

Machine-learning algorithm 'random forest' [89] from the software Weka [90] was employed to incorporate all features to train the classifier that best distinguishes false positive SNVs from true positive ones. The Random forest algorithm is employed with 100 trees, maximum $log2(n_{feature}) + 1$ features ($n_{feature}$ is the total number of features) to consider in each tree and no limitation on depth of the trees. The output of the classifier is a probability that a candidate SNV being a true SNV. The threshold is 0.5, thus if classification probability is greater than 0.5 then the candidate SNV is considered to be a true SNV.

2.2.6 Performance Evaluation

The AmpliSeq Comprehensive Cancer Panel targets exonic regions of known cancer related genes. Exonic loci with at least 5 reads supporting an alternative allele are included in the evaluation (Table 4). Precision and recall are defined in Equations (4) and (5). The allele frequency ranges were determined by the observed values for precision and expected values from test benchmark for recall.

$$precision = \frac{recovered\ test\ benchmark\ SNVs}{predicted\ number\ of\ SNVs} \quad (4)$$

$$recall = \frac{recovered\ test\ benchmark\ SNVs}{expected\ number\ of\ test\ benchmark\ SNVs} \quad (5)$$

2.2.7 Parameter Customization for Existing Tools

Existing tools to be compared with include TVC from Torrent Suite software, designed for Ion Proton sequencing data, Strelka and VarScan2.

TVC version 4.4-8 from Torrent Suite version 4.4.2: customized parameter setting was used since the default setting from TVC (Table 5) excludes SNV candidates with less than 3.5% (parameter gen-min-alt-allele-freq [91]) allele frequency. The minimal mapping quality for a read to be considered parameter 'MAPQ' was set to be the same as RareVar. There is no option for turning off 'downsample', thus the maximum depth (34,223) in test benchmark data was used.

Strelka [72] version v1.0.14: parameter file for bwa aligner was used. Depth filter on high-depth loci (isSkipDepthFilters [92]) was turned off. Also, since low recall was observed for <= 3% SNVs, combinations of 'ssnvPrior' and 'ssnvNoise' were tested. 'ssnvPrior' specifies the prior probability of a locus containing somatic SNVs while 'ssnvNoise' specifies the prior probability of a locus containing sequencing noise. The conclusion from this combinatory exploration suggested elevated 'ssnvNoise' decreases precision and recall while elevated 'ssnvPrior' increases recall with a slight drop in precision. The 1000x bigger 'ssnvPrior' results in ~ 3% increase in recall and ~ 1% drop in precision and since the extent of change is small, no further increase was attempted.

VarScan [71] version v2.3.7: the parameter for minimal SNV allele frequency ('min-var-freq' [93]) was set to 0.5% and minimal number of reads supporting alternative allele ('min-reads2') was set to 5 to be consistent with RareVar. Since the percentage of DNA from test benchmark 'normal' sample individual was 0.46, thus the parameter specifying the percent of tumor cell population ('tumor-purity') was set to 0.54. The complete list of parameters is in Table 5.

Table 4 RareVar benchmark results: number of somatic SNVs by frequencies.

	Training Benchmark			Testing Benchmark		
AF	SNVs	Training	UR[1] Rate	SNVs	Testing	UR Rate
0.5%	394	304	22.84%	388	270	30.41%
1%	319	271	15.05%	389	309	20.57%
1.5 to 3%	414	360	13.04%	569	493	13.36%
3.5 to 5%	185	162	12.43%	189	164	13.23%
5.5 to 10%	227	213	6.17%	161	151	6.21%
10.5 to 53%	170	151	11.18%	186	170	8.60%
All	1709	1461	14.51%	1882	1557	17.27%

1: UR stands for under-represented. A benchmark SNV is considered under-represented if fewer than 5 reads supporting the alternative allele.

Table 5 Adjusted parameters for tools compared with RareVar.

Parameter	Definition[1]	Default Value	Customized Value	Tool Name
snp-min-allele-freq	Minimum observed allele frequency required for a non-reference variant call.	0.02	0.005	TVC
gen-min-alt-allele-freq	Filter out variant candidates that do not have at least this frequency.	0.035	0.0025	TVC
MAPQ	Minimum mapping quality.	4	40	TVC
downsample	Reduce coverage in high-depth locations to this value.	2,000	34,223	TVC
isSkipDepthFilters	Binary tag to filter loci with high depth. 1 means no filtering.	0	1	Strelka
ssnvPrior	Prior probability of a locus contains somatic SNV.	1.00E-06	1.00E-03	Strelka
ssnvNoise	Prior probability of a locus contains sequencing noise.	5.00E-07	5.00E-07	Strelka
min-var-freq	Minimal SNV allele frequency.	0.2	0.005	VarScan
tumor-purity	Percent of tumor cell population.	1	0.54	VarScan
min-reads2	Minimal number of reads supporting the alternative allele.	2	5	VarScan

1: Definitions were adapted from documents for each tool.

2.3 Results

2.3.1 Benchmark Data Evaluation

After filtering reads with mapping quality less than 40, about 79 and 68 million reads were used for training benchmark normal and mixed tumor samples, respectively, and about 59 and 64 million reads were used for testing benchmark normal and tumor samples, respectively. The design of benchmark sets generates 1,709 and 1,882 somatic SNVs in the training and testing benchmarks, respectively. The design also ensures evaluation of SNVs with a broad range of allele frequencies, with special attention to the low frequency (0.5-3%) SNVs. As shown in Table 4, the percent of somatic SNVs with allele frequency no more than 3% is 65.9% in training benchmark and 71.5% in testing benchmark. Somatic SNVs with fewer than 5 reads supporting alternative allele were considered under-represented, and were excluded from training and testing dataset.

We checked the allele frequency agreement between sequenced benchmark tumor samples and the design. Potential SNV allele frequency bias introduced by pipetting variation in the pooling step was evaluated by the correlation of the detected median allele frequencies of SNVs unique to each individual with their designed frequencies (Table 1). The log scale linear regression analysis showed the individuals with smaller assigned percentages tend to have a slightly lower than the design percentages, with $R^2 > 0.98$ for both training and testing benchmarks (Figure 3). Thus the observed allele frequencies highly correlated with the design, we used the benchmark datasets for further modeling building and performance evaluation.

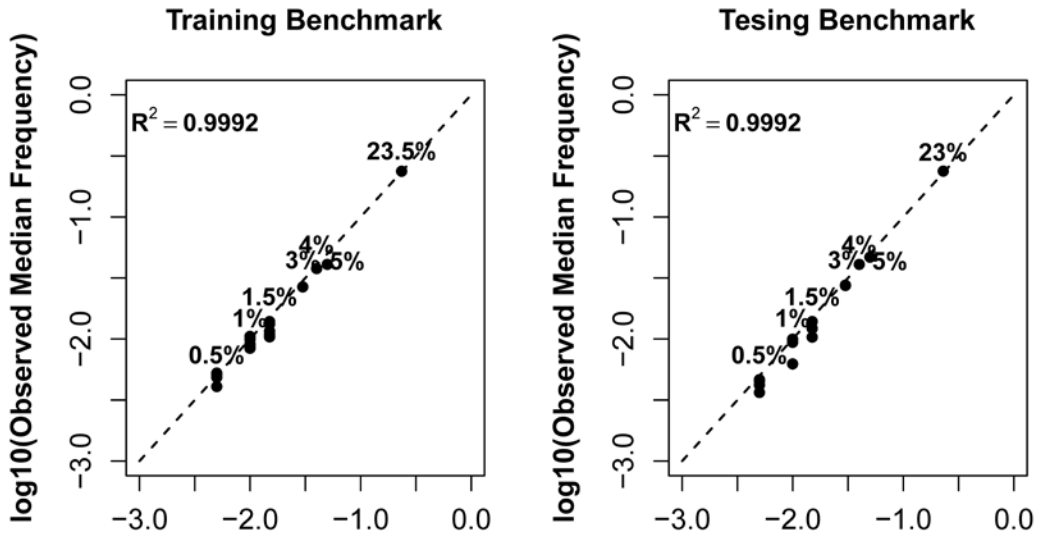

Figure 3 Evaluation of pipetting variance in construction of the training and testing benchmarks. Numbers next to the dots represent the mixing frequencies of DNA samples; the line at 45 degrees represents perfect pipetting (observed frequency exactly equals the expected). The R^2 for both training and testing benchmarks are from the linear regression results of using observed frequency in log10 scale as response variable – denoted by y and mixing frequency in log10 scale as explanatory variable – denoted by x. The coefficients for the explanatory variable are 1.025 (y = 1.025x) and 1.030 (y = 1.030x) for training and testing benchmarks, respectively.

2.3.2 Position Specific Error Model and Variant Identification

For each nucleotide different from the reference in each locus, a Poisson distribution generalized linear model (PD-GLM) was used to model the strand-specific sequencing error rate based on the associated genomic features (Table 2). Regression results showed all 9 features tested were statistically significant (p values < 0.001) in contributing to sequencing errors. Overall, the fitted model showed significant improvement compared to an intercept-only model (with no features considered), with Cragg & Uhler's [94] R^2 = 0.246, indicating an excellent fit for R^2 between 0.2 ~ 0.4 [95, 96]. The signs and relative magnitudes of coefficients agreed with prior knowledge and the intuition derived from visualization. Take the feature substitution as an example, clear transitional bias, or purine/pyrimidine conservation was supported by the PD-GLM since four substitution types (A > G, C > T, G > A, T > C) have the largest positive coefficients (Table 2). As for the neighbor nucleotide composition complexity effect, slightly increasing error rates were observed for higher GC content values (Figure 4 top left) and this observation is also reflected in the small positive coefficient value (β_{GC}= 0.005). Also, if the alternative base is the same as the immediate upstream or downstream (for example a trinucleotide pattern CTG, the center base T changes to either upstream letter C or downstream letter G), these alternative bases observed are more likely to be context-induced errors ($\beta_{alt\_up\_down}$ = 0.690). For homopolymer related features, a locus is more erroneous if it is within 2 nucleotides of more long homopolymer(s) (β_{hmer_dist} = -0.009, β_{hmer_len} = 0.081, $\beta_{hmer\_percent}$ = 0.193), and within a homopolymer (hmer_op = 0.375) (Figure 4 top right and bottom subplots). The magnitudes of the above-mentioned covariates indicate (1) if a candidate SNV locus is 1 nucleotide further from a neighbor homopolymer then the error rate in natural log scale drops by 0.009, (2) if the neighbor homopolymer length increases by 1 nucleotide then

the error rate in natural log scale increases by 0.081, (3) if the percent of homopolymer bases increases by 0.1 (range 0 to 1) then the error rate in natural log scale increases by 0.193*0.1 = 0.0193 and (4) if the locus is within a homopolymer then the error rate in natural log scale increases by 0.375.

After fitting the PSEM using PD-GLM, a Bayes factor was calculated for each base in each locus to determine its likelihood ratio of being a somatic SNV (from model M_V) rather than a sequencing error (from model M_E). To evaluate the efficacy of the PSEM in identifying SNV candidates, we mainly compared the performance of PSEM with Fisher's Exact Test based VarScan2 [71]. VarScan2 was picked for comparison here because different from other tools (Strelka and TVC) that utilized sequencing quality features to boost precision at the cost of reduced recall, VarScan2 only considers the depth and number of reads supporting the alternative allele at the same genomic position of the two samples being compared. Thus the underlying features utilized by VarScan2 are genomic sequence contexts determined at each locus, which are the features that PSEM utilized explicitly. Therefore, both VarScan2 and PSEM should be able to recover the most number of true SNVs, which will be reflected in higher recall, especially in lower frequency ranges. In Table 6A comparing recall of different tools, we did observe PSEM and VarScan2 standing out as top 2 tools in overall recall and also showing large advantages over Strelka and TVC in 0.5% to 3% allele frequency ranges. Further, compared with VarScan2, PSEM showed higher overall recall (95.8% versus 83.0%), with the advantages more evident at 0.5% (22.6% increase) and 1% (15.2% increase) ranges. However, comparing precision in Table 6B PSEM and VarScan2 showed lowest precision, especially at in 0.5% to 3% allele frequency ranges. Thus, despite the high efficiency of PSEM in recovering candidate SNVs, sequencing related features needed to be incorporated into the SNV caller for candidate recalibration.

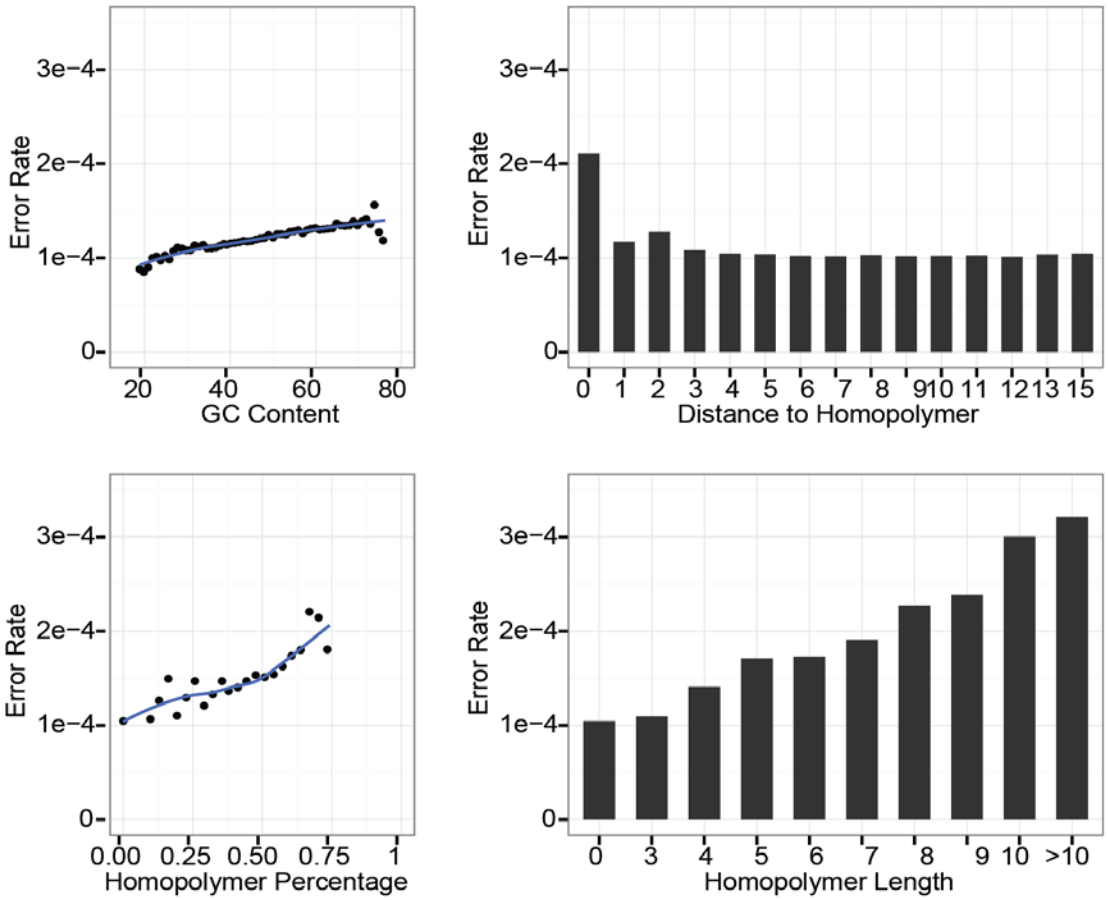

Figure 4 Relationship between genomic context features and error rate. The error rate is the mean of error rates from all data with a certain feature value, for example homopolymer length 7. For feature values with less than 1000 points, the points are combined to derive the mean error rate. The line is the smoothed trend line. The details of the features are described in Table 2.

Table 6 RareVar: Comparison of recall and precision for different allele frequencies.

A: Recall comparison

	VarScan2[2]			PSEM[3]		
Expected Frequency [1]	**Expected Number of SNVs**	**Recovered Number of SNVs**	**Recall**	**Expected Number of SNVs**	**Recovered Number of SNVs**	**Recall**
0.5%	270	169	62.6%	270	230	85.2%
1%	309	249	80.6%	309	296	95.8%
1.5 to 3%	493	438	88.8%	493	483	98.0%
3.5 to 5%	164	148	90.2%	164	162	98.8%
5.5 to 10%	151	137	90.7%	151	150	99.3%
10.5 to 54%	170	152	89.4%	170	170	100.0%
All	1557	1293	83.0%	1557	1491	95.8%
	Strelka[2]			**RareVar[4]**		
Expected Frequency [1]	**Expected Number of SNVs**	**Recovered Number of SNVs**	**Recall**	**Expected Number of SNVs**	**Recovered Number of SNVs**	**Recall**
0.5%	270	0	0.0%	270	96	35.6%
1%	309	0	0.0%	309	251	81.2%
1.5 to 3%	493	121	24.5%	493	458	92.9%
3.5 to 5%	164	138	84.1%	164	156	95.1%
5.5 to 10%	151	141	93.4%	151	149	98.7%
10.5 to 54%	170	160	94.1%	170	165	97.1%
All	1557	560	36.0%	1557	1275	81.9%
	TVC[2]					
Expected Frequency [1]	**Expected Number of SNVs**	**Recovered Number of SNVs**	**Recall**			
0.5%	270	7	2.6%			
1%	309	88	28.5%			
1.5 to 3%	493	429	87.0%			
3.5 to 5%	164	158	96.3%			
5.5 to 10%	151	146	96.7%			
10.5 to 54%	170	166	97.6%			
All	1557	994	63.8%			

1: SNV frequencies are based upon the mixing scheme in Table 1.

2: The customized parameters applied are listed in Table 5.

3: PSEM represents the intermediate results of RareVar framework candidate SNV calling step using Bayes factor.

4: RareVar contains both PSEM and machine-learning base recalibration steps.

B: Precision comparison

	VarScan2[2]			PSEM[3]		
Expected Frequency[1]	**Predicted Number of SNVs**	**Recovered Number of SNVs**	**Precision**	**Predicted Number of SNVs**	**Recovered Number of SNVs**	**Precision**
0.25 to 0.75%	1105	223	20.2%	1369	286	20.9%
0.75 to 1.25%	389	256	65.8%	479	298	62.2%
1.25 to 3%	449	368	82.0%	571	422	73.9%
3 to 5%	175	158	90.3%	217	168	77.4%
5 to 10%	158	144	91.1%	207	159	76.8%
10 to 54%	149	144	96.6%	190	158	83.2%
All	2425	1293	53.3%	3033	1491	49.2%
	Strelka[2]			**RareVar[4]**		
Expected Frequency[1]	**Predicted Number of SNVs**	**Recovered Number of SNVs**	**Precision**	**Predicted Number of SNVs**	**Recovered Number of SNVs**	**Precision**
0.25 to 0.75%	0	0	NA	140	127	90.7%
0.75 to 1.25%	0	0	NA	271	264	97.4%
1.25 to 3%	89	84	94.4%	417	403	96.6%
3 to 5%	190	167	87.9%	172	165	95.9%
5 to 10%	182	157	86.3%	169	159	94.1%
10 to 54%	162	152	93.8%	166	157	94.6%
All	623	560	89.9%	1335	1275	95.5%
	TVC[2]					
Expected Frequency[1]	**Predicted Number of SNVs**	**Recovered Number of SNVs**	**Precision**			
0.25 to 0.75%	1	0	0.0%			
0.75 to 1.25%	104	97	93.3%			
1.25 to 3%	454	420	92.5%			
3 to 5%	194	183	94.3%			
5 to 10%	189	176	93.1%			
10 to 54%	170	163	95.9%			
All	1112	1039	93.4%			

1: SNV frequencies are based upon the percent of reads supporting alternative allele from sequencing.
2: The customized parameters applied are listed in Table 5.
3: PSEM represents the intermediate results of RareVar framework candidate SNV calling step using Bayes factor.
4: RareVar contains both PSEM and machine-learning base recalibration steps.

2.3.3 Machine Learning Based SNVs Calibration

In order to further reduce the number of false positive SNVs identified using the PSEM, we utilized a machine-learning model – Random Forest [89] to better distinguish true positive and false positive SNVs. We used information gain to rank the classification power of 29 measurements related to sequencing technology and downstream analysis methods together with all features from PSEM. Sequencing-related features and alignment quality features ranked the highest, while GC content was removed due to 0 information gain (Table 3).

The machine-learning-based variant recalibration effectively reduced false positive SNVs identified by the PSEM alone. In the testing benchmark dataset, the overall precision increased from 49.2% (after the PSEM model) to 95.5% (after machine-learning refinement, represented as RareVar in Table 6), with the overall recall rate dropped from 95.8% to 81.9%.

2.3.3.1 Performance by allele frequencies

A closer examination of the RareVar performance by allele frequencies showed the precision increased for all allele frequencies by at least 10% after machine-learning-based variant recalibration (represented as RareVar in Table 6) compared with PSEM alone, with greater than 90% precision achieved for SNVs in all allele frequency ranges (Table 6B). As expected, lower frequencies showed higher increase, in which 0.70 and 0.35 increases in precision were achieved for 0.5% and 1%, resulting in 90.7% and 97.4% precision respectively. The decrease in recall was mainly attributed to 0.5% and 1%, yet > 80% recall was maintained for allele frequencies ≥ 1%. The ROC curve on SNVs of different allele frequency ranges (Figure 5a) showed the model reaches relatively stable performance for SNVs with greater than 1% frequency.

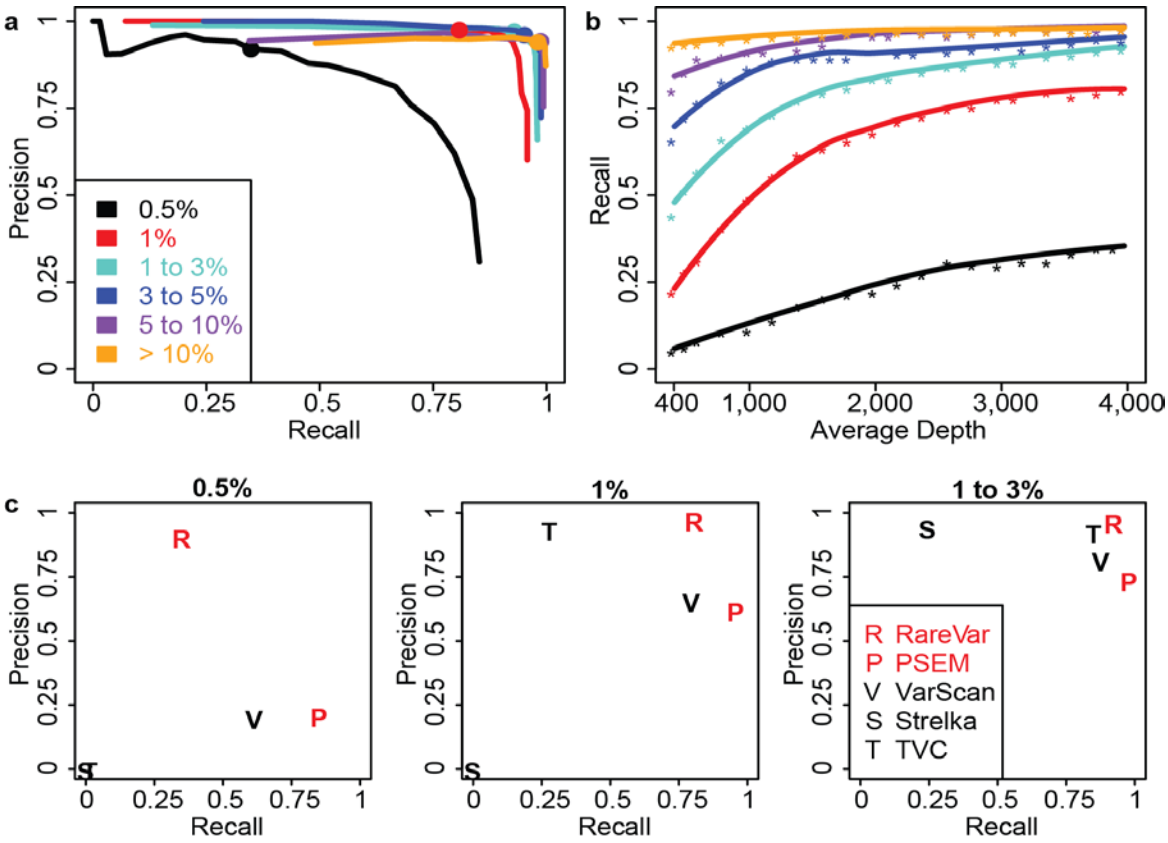

Figure 5 RareVar: Performance evaluations and comparisons. a. The precision and recall are evaluated at classification probabilities from 0 to 0.95, in steps of 0.05. Points with 0.50 probability are highlighted. The classification probability is outputted by the machine-learning algorithm, which evaluates the probability of a candidate SNV being a true SNV. 0.50 is the threshold. b. The depth is sampled from 10% to 100% of the original, in steps of 5%. c. Benchmark performance optimized parameters (Supplementary table 6) were applied for VarScan2, Strelka and TVC to compare with PSEM and with RareVar.

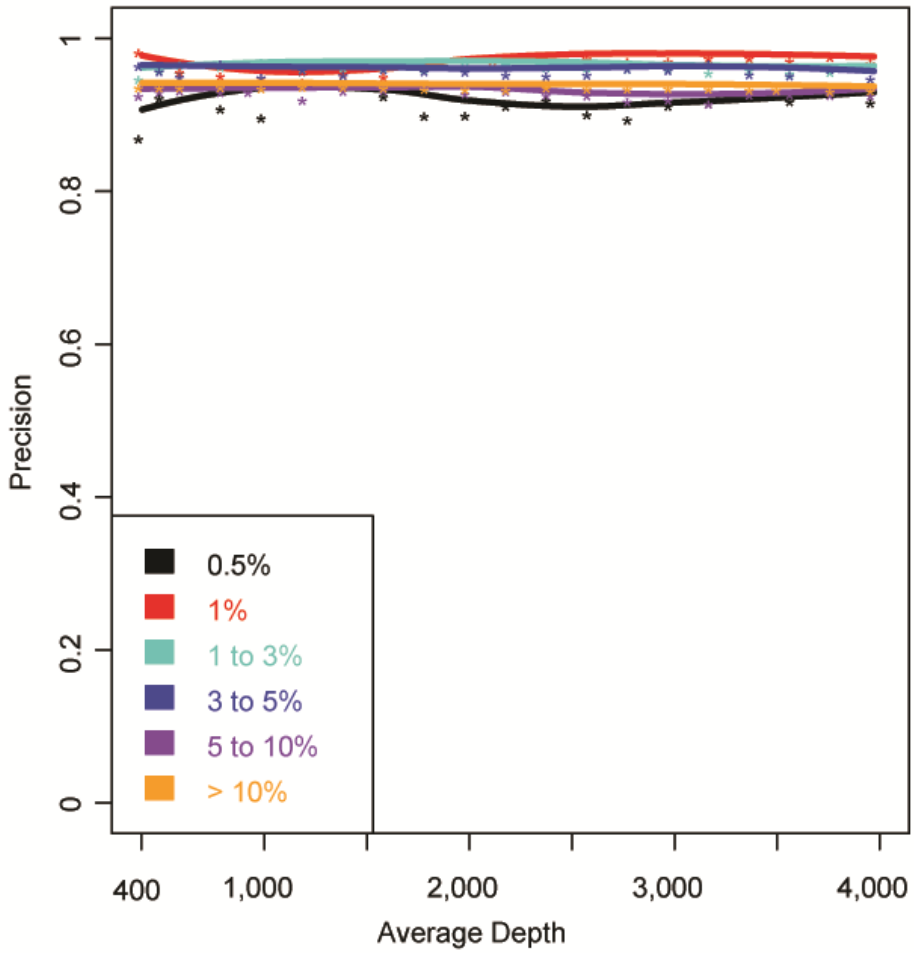

Figure 6 RareVar: comparison of precision at various frequencies.

2.3.3.2 Effects of Sequencing Depth on the Model Performances

Another factor affecting variant identification is the sequencing depth. The average sequencing depth for the testing set tumor sample (Table 1) was 3,973. In order to evaluate the influence of sequencing depth on the precision and recall for the SNVs at various allele frequencies, we gradually down-sampled the testing benchmark tumor sample sequencing data by randomly selecting a fixed percentage of reads from the original sequencing data. The precision is not affected (Figure 6), but the recall steadily decreases with reducing average depths for SNVs at all ranges of frequencies (Figure 5b). It is also obvious that the trend for the decreasing is more severe when the sequencing depth is less than 1000x. This result suggests that sequencing depth should be pre-determined for detecting SNVs at a specific frequency range. In addition, for the variants whose allele frequency is greater than 0.5%, the recall curve reaches a plateau when average sequencing depth is greater than 2000x. This suggests that further increasing the sequencing depth won't improve the sensitivity of the detection.

2.3.4 Performance Comparison with Existing Methods

We compared RareVar with established variant detection tools including VarScan2, TVC from Torrent Suite software, and Strelka. Since the default settings for these tools aim at SNVs with higher allele frequency, customized parameters were selected by optimizing for rare variants (details in Method and Table 5). RareVar was the best method in overall performance (Figure 5c and Table 6), with the advantage over the second best method, TVC, most evident for 1% and 0.5% frequencies. At 1% allele frequency, the precision is similar for both methods, while RareVar achieved 81.2% recall, compared with 28.5% for TVC. Even at 0.5% allele frequency, RareVar maintained 90.7% precision and 35.6% recall. It is expected that TVC also demonstrates

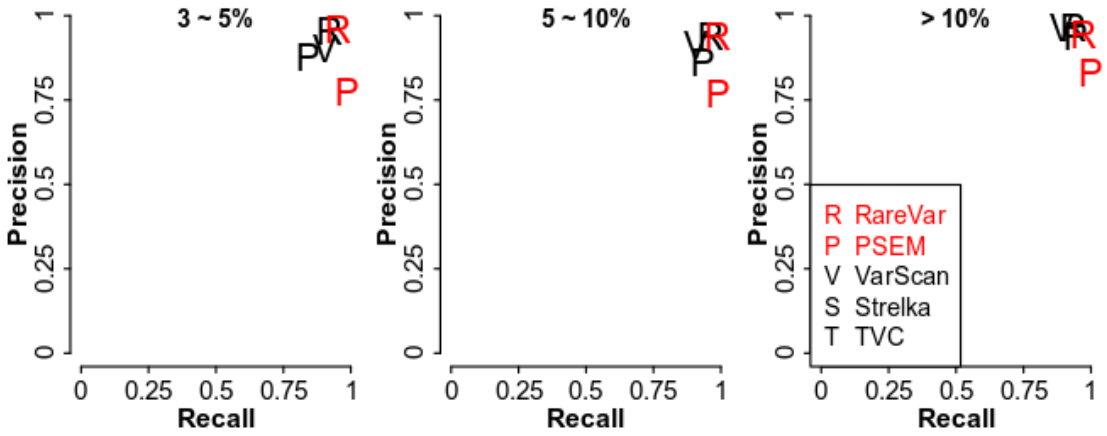

Figure 7 RareVar: comparison of precision and recall for common SNVs.

good performance for > 1% since it is specifically designed for Ion Proton technology, by using post-alignment correction for homopolymers. Although Strelka achieved comparable performance for more frequent SNVs (≥10%) (Figure 7), it showed unsatisfactory results for SNVs with allele frequency < 3%. Overall, RareVar shows the best performance among all the tools tested, in particular for the SNVs with low allele frequencies (0.5%-3%).

2.4 Discussion

The framework of RareVar provides guidance for low frequent SNV identification from both experimental and algorithmic aspects. Many components in the next generation sequencing pipeline, including library preparation, target enrichment assay and sequencing technology, affect the sensitivity and fidelity of SNV detection. The comparison of RareVar with other algorithms underscores the necessity of modeling frequency detection limits and the significance of a model tailored for each technology. It is impractical to have a universal parameter or threshold setting scheme that fits all sequencing platforms and experimental protocols. To solve this problem, we construct a benchmark sample containing variants with desired allele frequencies. The distribution of the nucleotide mismatch patterns around the positive and negative variant loci in the benchmark sample provides a valuable guideline for optimizing the parameter and threshold settings during the variant identification process. In addition, the independent testing benchmark samples also enables fair evaluation on the performance of the detection.

The two stages in the computational modeling, position-specific error model (PSEM) and machine-learning-based recalibration, were designed to take the advantages of sequencing signals on the invariant loci, and designed variant loci, respectively. The PSEM step intends to model how genomic sequence contexts impact

the sequencing error profiles that are associated with the experimental protocol. The derived model serves as an important base for accurately estimating background error signal that are specific to any particular nucleotide position. This step is critical in improving the detection accuracy of the SNVs with extreme low frequency, as opposed to using a universal background error rate for all the genomic loci.

A machine-learning-based algorithm is applied in the variant recalibration stage, in which experiment-related features, such as strand bias and mapping quality, are considered. This design effectively avoids using a series of filters that often involves multiple ad hoc thresholds. Conceptually, this is similar to the variant recalibration strategy used in GATK with two major differences. First, GATK assumes that identified variant loci documented in the dbSNP database are likely to be true positive variants. This assumption is not valid for cancer somatic mutations, in which most mutation loci are random and the allele frequencies in the sample vary. Second, GATK constructed a Bayes Gaussian mixture model only based on true positive variant loci while false positive loci are difficult to determine without a gold standard dataset. With our strategy, however, both true positive and true negative variants are available for the benchmark sample. This enables using more sophisticated machine learning algorithm, such as random forest model. We demonstrated that variant recalibration step significantly increase the specificity of the variant identification, and further improved the overall accuracy.

Chapter 3. Analyzing Mutational Drift and/or Enrichment in Reprogrammed Primary Breast Tumor Cells

3.1 Introduction

Reliably detecting low frequency SNVs from NGS sequencing data provides great promises for cancer research and clinical applications. Deep sequencing targeted regions to measure the low frequency SNVs is a commonly used approach [15, 97, 98]. However, NGS sequencing platform artifact level (0.1~1%) puts a barrier in terms of how low frequency can be detected. In addition, intratumor heterogeneity is prevalent in tumors [99, 100]. Thus directly sequencing bulk tumors might not allow us to detect low-prevalent subpopulations with actionable mutations. Multiregion sequencing was employed to uncover the intratumor heterogeneity by sequencing multiple spatially different biopsies and checking the geographically distinct patterns of somatic mutations [101]. Yates et al [102] sequenced breast tumor biopsies from multiregion and confirmed the importance of considering subclonal structure in breast cancer research and clinical trials. Another way of considering subclonal structure is to provide suitable conditions and allow the subclones to grow, and thus the minor subpopulations with growth advantage may grow and be detectable. Comparing the mutational profiles of cultivated tumor cells with that from the original tumor tissue, the mutational drift and/or enrichment in cultivated tumor cells could be revealed. These events provide insights about the tumor subpopulation evolution as well as clues for cancer treatment.

To cultivate tumor cells, cell reprogramming was used to induce the tumor tissue cells to grow indefinitely in vitro [103]. The reprogramming step cultivates primary tissue cells using irradiated mouse embryonic fibroblasts as feeders and media containing ROCK inhibitor [104, 105]. To detect somatic mutations from unprocessed primary tumor cells and potential mutational drift and/or enrichment from reprogrammed tumor, we sequenced DNA samples from the following cell types: patient peripheral blood,

unprocessed tumor tissue and tumor adjacent normal tissue, reprogrammed tumor cells and tumor adjacent normal cells. The mutations from peripheral blood represent the germline mutational landscape and thus can be used to determine somatic mutations in tumor samples. In the absence of blood samples, unprocessed tumor adjacent normal tissue cells are used to derive germline mutations. The unprocessed tumor tissue cells contain the somatic mutations from multiple tumor cell subpopulations, in which some populations may harbor malignant mutations that bear growth advantage over others. However, if sequencing the bulk tumor, some of the minor tumor subpopulations with unique mutations may not be detectable. We hypothesized that the reprogrammed cells allow the expansion of the minor population of tumor cells with growth advantage. Thus, by comparing primary tumor cell mutational profile with that from its reprogrammed counterpart, we can identify mutational drift and/or enrichment. The reprogrammed cells may contain mouse cells, which could be mistaken as human somatic mutations if not carefully characterized. We applied bioinformatics approaches to filter sequencing reads likely from mouse cells.

3.2 Materials and Methods

3.2.1 Materials and Sequencing

Blood, fresh adjacent normal and tumor tissues from 5 patients were obtained from Indiana University Simon Cancer Center (IUSCC) Tissue Bank. Table 7 summarized the number of samples sequenced from each type, the patients are grouped as ‘A’ and ‘B’ depending on the available sample types. Among these patients 1402-17 is unique in that she had different types of breast tumors.

The collaborators prepared the reprogrammed tumor cells. Briefly, the tissue samples from IUSCC were split into two. One part was frozen and the other part was used to cultivate primary cells using irradiated mouse embryonic fibroblasts as feeders

and media containing ROCK inhibitor [105]. Jam-A/EpCAM were used to remove mouse fibroblasts. Jam-A+/EpCAM+ cells were sorted by flow cytometry and thus mouse fibroblasts were removed since they did not stain for these markers.

All samples were sequenced with Ion Proton. The sequencing library preparation and sequencing data alignment as well as post-alignment filtering were the same as described in Chapter 2 section 2.2.2.

Table 7 Summary of breast tumor samples and types sequenced.

		Unprocessed		Reprogrammed	
Patient ID	**Group**	**Blood / Normal**	**Tumor**	**Normal Cell**	**Tumor Cell**
1406-26	B	Adjacent Normal	Yes	Yes	Yes
1411-04	B	Adjacent Normal	Yes	Yes	Yes
1310-33	A	Blood	Yes		Yes
DCIS	A	Blood	Yes		Yes
LCIS	A		Yes		Yes

3.2.2 In-silico Characterizing and Filtering Sequencing Reads Originating from Mouse

To characterize the degree of contamination from mouse cells, we first applied in silico PCR to tabulate the percentage of amplicon primer pairs that can also specifically pull down sequences from mouse genome. The in silico PCR tool from UCSC genome browser [106] is used to search over ~16,000 amplicon primer pairs from the Ion AmpliSeq Comprehensive Cancer panel. The default parameters for in silico PCR tool in UCSC genome browser were used, which required 15bp perfect match for both 5' and 3' primers and also maximum 4000bp amplified region. We found that 235 primer pairs can also pull down mouse genome sequences, which is 1.47% of all amplicons.

Despite the low percentage of amplicon primer pairs that may introduce mouse DNA contamination, it is still necessary to consider the possibility that the primer pairs may have some level of random pairing which may potentially pull down mouse genome sequences. However, the combinatory search space is huge (16,000 * 16,000), thus we explored several methods of finding mouse genome reads based on comparative alignment between the mouse and human genome.

The sequencing reads derived for cultured reprogrammed cells were mapped both to the human genome (genome build hg19) and the mouse genome (genome build mm10) [107] using TMAP from Torrent Suite software. To distinguish reads from mouse rather than human, we explored 3 strategies to filter the sequencing data, (1) 'no mouse' which removes reads that can be aligned to the mouse genome with mapping quality greater than 20, (2) 'MAPQ' which removes reads that have a larger mapping qualities when mapped to the mouse genome than the human genome, (3) 'longer match' which removes reads that have a larger total number of aligned bases to the mouse genome than the human genome.

To compare the performances of the filtering methods, different strategies were designed for the two patient groups due to the different availability of sample types in the two groups. For group A, to evaluate the consequences of false positive filtering, we checked the agreement of SNVs between unprocessed tumor and after applying mouse read filtering. Since there should be no mouse cell contamination in unprocessed tumors, any reads removed are false positive mouse reads. For group B, to evaluate the efficiency of different methods in removing reads from mouse cells, we checked the agreement of SNVs between unprocessed normal tissues and reprogrammed normal cells. Since there should be no new mutations or small number of mutations potentially induced by reprogramming process in normal cells, any new SNVs from normal cells compared with unprocessed normal tissues were considered as false positive SNVs due to contamination from mouse cells. We calculated recall, precision, and F1 score as described in section 3.2.3.

3.2.3 Detecting Somatic SNVs with RareVar

SNV detection was done with RareVar described in chapter 2, to effectively deal with diluted SNV signals from low-prevalence tumor subpopulations. For each patient, all types of samples independently went through the Bayes factor based candidate SNV identification and machine-learning based recalibration in RareVar framework to derive SNVs, then we applied a series of filters and statistical tests to determine somatic SNVs. Step 1: filtered candidate somatic SNVs by only including SNVs (1) not in potentially mouse contaminated amplicons, (2) RareVar detected those SNVs in either tumor tissue or tumor cells and the allele frequencies are larger than those in the germline sample, (3) depths on SNV loci in tumor tissue and tumor cells are greater than 100 and (4) maximum of allele frequencies from tumor tissue and tumor cells are at least 2-fold of the allele frequencies from germline sample. Step 2: for SNVs detected in both tumor

tissue and tumor cells, a binomial test (p value threshold 0.01, single sided test) was first used to check if the allele frequencies are significantly larger than those in germline sample. Then only the ones showing larger frequencies were kept and went through a second binomial test to see if the allele frequencies in tumor tissue are different from those in tumor cells. If allele frequencies are significantly (p value threshold 0.01, single sided test) greater in tumor cells, then those SNVs potentially are from enriched tumor subpopulations in tumor cells. If the allele frequencies in tumor cells are smaller or similar, then the prevalence of these host tumor subpopulations possibly did not change. Step 3: for SNVs only detected in tumor tissue by RareVar, we first used binomial test to make sure the allele frequencies were greater than those in the germline sample, then checked whether there are also reads supporting those SNVs in tumor cells. If there are, it is an indicator of the host subpopulation shrinkage (the percentage in tumor cells is smaller than in tumor tissue) and also increases our confidence that those are true somatic SNVs rather than sequencing artifacts. Step 4: for SNVs only detected in tumor cells by RareVar, we first used binomial test to make sure the allele frequencies were greater than those in the germline sample, then checked whether there are also reads supporting those SNVs in tumor tissue. If there are, it is an indicator of the host subpopulation enrichment and also increased our confidence that those are true new somatic SNVs rather than sequencing artifacts.

3.3 Results

3.3.1 Removing Contaminating Reads from Residual Mouse Cells

We first explored the percentage of reads removed by all methods described in section 4.3.1. Taking 1406-26 tumor cell sample as an example, ~32% reads were removed by ‘no mouse’ method while only 1~2% reads were removed by ‘MAPQ’ or ‘longer match’ method. This result agrees with the speculation that ‘no mouse’ method

tends to remove many reads from human and mouse homologous regions, since the protein-coding regions of human and mouse genomes are on average 85 percent identical [109]. Thus, we further explored the effectiveness of the other two methods by comparing the consistency of detected SNVs from tumor/normal cell with unprocessed tumor/normal tissue samples.

We checked the effect of falsely removing reads from human on SNV calling in group A. The key values are the number of SNVs after read filtering that overlapped with unfiltered samples, referred as 'Overlapped with UP Tumor' in Table 8A. For all three samples, 'longer match' and 'MAPQ' performed similarly. 'MAPQ' correctly recovered 1 more SNV from 1310-33 while 'longer match' correctly recovered 4 more SNVs from DCIS (Table 8A). When visually comparing the performances of the two methods using the recall, precision and F1 score measures, no visible differences could be observed except for the recall for DCIS (Figure 8).

We checked the effect of failing to remove reads from mouse cells on SNV calling in group B. The key values are the number of SNVs after read filtering that overlapped with unprocessed normal tissue samples, referred as 'Overlapped with UP Normal' in Table 8B. The numbers of overlapped SNVs were slightly higher in 'longer match' for both samples. Besides, 'MAPQ' filtered data had more SNVs identified, 57 more in 1406-26 and 804 more in 1411-04. We hypothesized there should be no or only small number of new SNVs in reprogrammed normal cells, thus the 'MAPQ' method is considered to be less efficient in removing reads from mouse cells. When visually comparing the performances of the two methods using the recall, precision and F1 score measures, visible differences could be observed for the recall and F1 score for both samples in group B (Figure 8). Thus, 'longer match' was used as the read filtering method.

Table 8 Comparing methods for removing reads from mouse cells. Samples with blood, unprocessed tumor tissue and reprogrammed tumor cells are grouped in A. Samples with unprocessed normal and tumor tissues, as well as reprogrammed normal and tumor cells are grouped in B. UP Tumor: unprocessed tumor. UP Normal: unprocessed adjacent normal tissue. Normal Cell: reprogrammed normal cells. For cells with a single number, that number is the number of SNVs detected. For cells with 2 numbers, the configuration is explained in the third point by the end of table B.

A

		MAPQ Filter[1]	**Longer Match Filter[2]**
Sample	**UP Tumor**	**Filtered UP Tumor / Overlapped with UP Tumor**	**Filtered UP Tumor / Overlapped with UP Tumor**
1310-33	1143	1143 / 1142[3]	1141 / 1141
DCIS	1150	1145 / 1145	1150 / 1149
LCIS	1139	1141 / 1138	1141 / 1138

B

		MAPQ Filter[1]	**Longer Match Filter[2]**
Sample	**UP Normal**	**Filtered Normal Cell / Overlapped with UP Normal**	**Filtered Normal Cell / Overlapped with UP Normal**
1406-26	1117	1192 / 1076	1135 / 1080
1411-04	1123	1977 / 1076	1173 / 1077

1: MAPQ Filter - remove reads that have a larger mapping quality when mapped to mouse genome.
2: Longer Match Filter - remove reads that have a larger total number of mapped bases when mapped to mouse genome.
3: In this example, 1143 is the number of SNVs detected in filtered UP tumor sample, while 1142 is the number of SNVs from filtered UP tumor sample that overlapped with UP tumor sample. The other cells with the format ‘number 1’ / ‘number 2’ could also be explained by checking their column names.

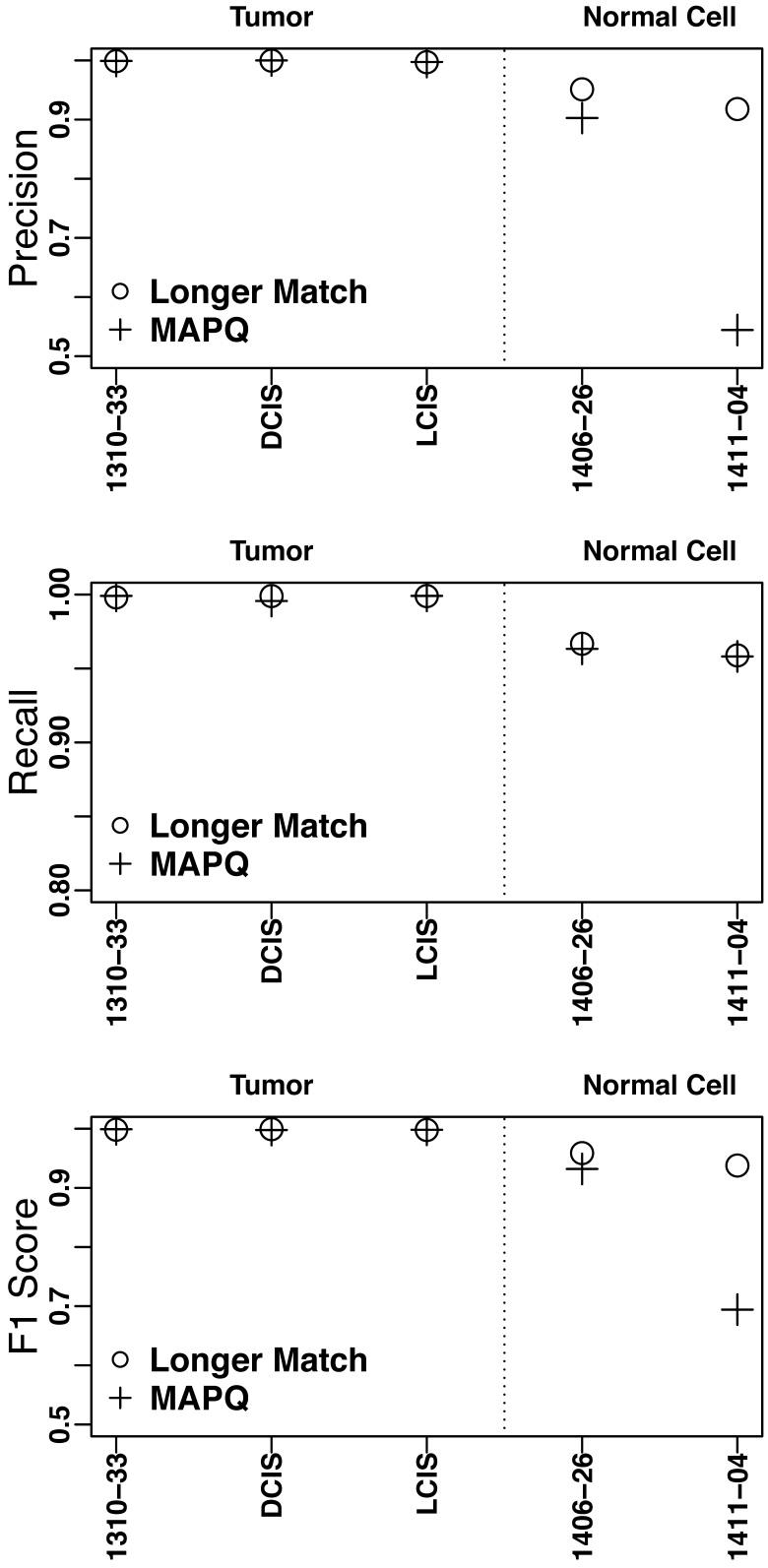

Figure 8 Comparing methods for removing reads from mouse cells.

3.3.2 Summary of Identified SNVs and Their Functional Implications

We derived the somatic SNVs based on the 4 steps applied to RareVar results described in section 3.2.3. A summary of the SNVs of different potential functional groups was list in Table 9. For all somatic candidate SNVs, they are mostly identified in either tumor tissue or tumor cells but not both. Number of SNVs possibly going through somatic enrichment is more than those of somatic shrinkage.

For every patient, we selected candidate SNVs for further examination. Reprogrammed tumor cells from patient 1310-33 showed splice site mutation of ETV1 oncogene [110-114], and both tumor tissue and cells showed a stop-gain mutation from tumor metastasis-associated gene MYH9 [115, 116]. In tumor cells from a patient (1402-17) with lobular carcinoma in situ (LCIS) in her left breast, we detected PI3KCA H1047R mutation, a common mutation found in many cancers [117-122]. In this tumor, we also detected R554C mutation of FOXO1, a context-dependent tumor suppressor or oncogene [123-128]. This is a novel mutation yet to be reported in any cancer. This patient had invasive carcinoma with ductal carcinoma in situ (DCIS) in her right breast. This tumor had multiple mutations in PIK3R1 gene (L7R, L100R, L70R, and L370R). Although PI3KR1 mutations are very common in cancer [129-132], these specific mutations have not been reported (as per cBioportal [133, 134]). Tumor cells from patient 1406-26 were detected to have a novel nonsynonymous SNV in FZR1 gene (N315S, N404S), which is a candidate CDK4/6-cyclin D substrate [135]. Tumor from patient 1411-04) showed mutation in the epigenetic regulator DAXX [136-139]. R371W mutation of DAXX has previously been reported in two cases of AML [140].

Table 9 Summary of detected somatic SNVs in breast tumor samples.

RareVar SNV Prediction		Alternative AF[3] Comparison	Possible Event[4]	Number of Somatic SNVs				
UP Tumor[1]	Tumor Cell[2]			1310-33	DCIS	LCIS	1406-26	1411-04
Yes	Yes	Higher in Tumor Cells	Somatic SNV Enrichment	5	0	0	0	0
Yes	Yes	Lower in Tumor Cells	Somatic SNV	0	1	0	0	0
Yes	Yes	Similar	Somatic SNV	3	0	0	0	0
Yes	No	Low in Tumor Cells	Somatic Shrinkage	3	10	2	4	2
Yes	No	0 in Tumor Cell	Somatic in Tissue	3	4	3	5	3
No	Yes	Low in UP Tumor	Somatic SNV Enrichment	9	7	102	4	5
No	Yes	0 in UP Tumor	New Somatic SNV in Cells	9	6	21	7	3

1: UP Tumor means unprocessed tumor.
2: Tumor cell means reprogrammed tumor cells.
3: AF: allele frequency. Comparing the allele frequencies in tumor tissue and tumor cells.
4: Biological events that possibly result in the observed allele frequency change.

3.4 Discussion

From experiment design side, the technique adapted in this study of sequencing unprocessed tumor and cultured cells from tumors will help to detect novel actionable mutations, which are otherwise missed by sequencing only bulk tumors. Variant caller designed specifically for sensitive low frequency mutation calling greatly facilitated the exploration of previously unknown genetic territories. Despite the novel findings, subsequent validation and functional characterization are indispensible to link SNVs to the functional disruption and thus uncover new drug targets.

Chapter 4. Statistical Modeling for Ion Proton Sequencing Platform Genomic Sequence Context Dependent Error

4.1 Introduction

In chapter 2, the RareVar framework for low-frequency single nucleotide variant detection was introduced. In RareVar, position specific error model (PSEM) using genome sequence context features is a key step. It is indispensable for determining lowest frequency detection limit as well as identifying candidate SNVs for downstream sequencing quality based candidate recalibration. Poisson distribution, a popular choice of distribution in modeling count date, was implemented under generalized linear model framework. However, the potential to improve PSEM performances on SNVs with close to sequencing error rates by implementing more sophisticated statistical distributions remains to be explored. In this chapter, we explored what distributions fit the DNA-Seq erroneous read count modeling as well as the possibility of improved position specific error rate prediction for higher precision and recall on SNVs down to 0.5% frequency. We reused the training and testing benchmark data sets sequenced with Ion Proton platform from chapter 2.

4.2 Materials and Methods

The focus of this chapter is to explore which statistical distribution fits the next generation sequencing error count data better. The general workflow for position specific error modeling using different distributions is described in Figure 9**.** In the training phase, starting from training benchmark normal sample invariant loci, the genome context features are extracted and fed to generalized linear models based on 4 candidate distributions. The genome context extraction and the fitted generalized linear models constitute the position specific error model. In the testing phase, all loci in the testing benchmark paired normal and tumor samples go through the position specific error

model and the candidate SNVs significantly different from fitted sequencing errors are generated. The following sections described in details the benchmarks and the configurations for generalized linear models based on different statistical distributions.

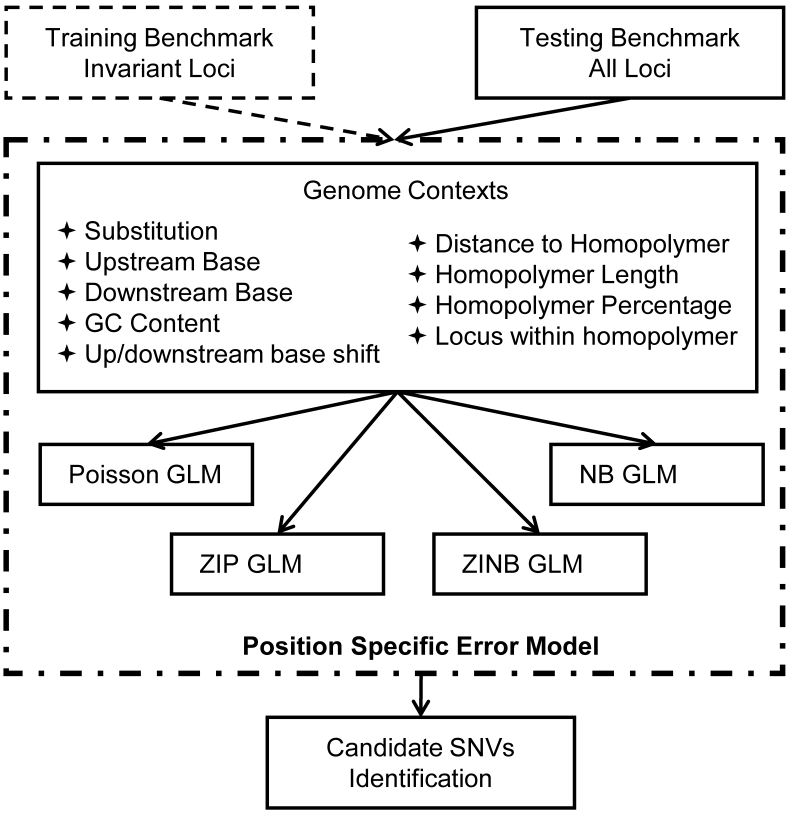

Figure 9 Diagram of the position specific error model using different statistical distributions. The dashed boxes highlighted the training data and the trained model. GLM: generalized linear model. NB: negative binomial. ZIP: zero-inflated Poisson. ZINB: zero-inflated negative binomial.

4.2.1 Benchmark Datasets

The Ion Proton training and testing benchmark datasets were generated as described in chapter 2 section 2.2.1.

4.2.2 Identifying Distribution Form for Sequencing Error Modeling

To model error rate based on count data, the 3 most common distribution choices are binomial, *Poisson* and negative binomial distributions. We applied a graphical exploratory plot – distplot [141-143] on the model response – number of reads containing non-reference bases – to get visual intuition about the overall fit of response data on different distributions. Intuitively, if an assumed distribution fits the data well, the data points should follow a straight line determined by the distribution metameters [141-143, 145]. The metameter of a discrete distribute equals a linear function of the count data (*k*), with the slope and intercept being the functions of distribution parameters. Under the context of sequencing error modeling, the count data *k* is observed number of reads supporting an alternative allele at a specific genomic locus.

Poisson distribution is taken as an example here to illustrate the form of the metameters. Assume a Poisson distribution with some fixed parameter mean λ, the observed frequency n_k for a value *k* equals the expected frequency $N * p_k$, where N is the total number of data points and p_k is the probability of observing *k*. Thus, setting $n_k = Np_k = Ne^{-\lambda}\lambda^k/k!$, and taking logs of both sides gives

$$\log(n_k) = \log(N) - \lambda + klog\lambda - logk! \quad (6)$$

This can be rearranged to a linear equation in *k*,

$$\emptyset(n_k) = log\left(\frac{k!\, n_k}{N}\right) = -\lambda + (log\lambda)k \quad (7)$$

The left side of equation is called the count metameter, and denoted $\emptyset(n_k)$. Hence, plotting $\emptyset(n_k)$ against k should give a straight line of the form $\emptyset(n_k) = a + bk$ with slope $log\lambda$ and intercept $-\lambda$, when the observed frequencies follow a Poisson distribution. The metameters slopes and intercepts for binomial and negative binomial distributions are summarized in Table 10.

Table 10 Distplot parameters for three discrete distributions. In each case the count metameter is plotted against k, yielding a straight line when the data follow the given distribution. k is the count data to be checked for appropriate distributions.

Distribution	Binomial	Poisson	Negative Binomial
Probability Function, p(k)	$\binom{n}{k} p^k (1-p)^{n-k}$	$e^{-\lambda}\lambda^k/k!$	$\binom{n+k-1}{k} p^n (1-p)^k$
Count Metameter, $\emptyset(\mathrm{n_k})$	$log\left(n_k/N\binom{n}{k}\right)$	$\log(k!\, n_k/N)$	$log\left(n_k/N\binom{n+k-1}{k}\right)$
Theoretical Slope	$\log\left(p/1-p\right)$	$\log(\lambda)$	$\log(1-p)$
Theoretical Intercept	$nlog(1-p)$	$-\lambda$	$\mathrm{n}\log(p)$

Table adapted from Hoaglin and Tukey (1985) [145], Table 9-15.

4.2.3 Generalized Linear Models

The details of the 9 genomic sequence contexts considered in generalized linear models were summarized in Table 2. These 9 features are the covariates included in the GLMs.

The Poisson distribution GLM specification remains the same as described in chapter 2 section 2.2.2. For the purpose of comparing with other distributions, the equation and variable descriptions are included below. The Poisson GLM for erroneous sequencing read counts with log link function is expressed in equation (8), where $N_{s,b,l}$ is the observed number of erroneous reads for strand *s* (forward or reverse) with alternative base *b* (three possible values other than the reference) at location *l*, $\lambda_{s,b,l}$ represents the expected mean for $N_{s,b,l}$, $\boldsymbol{c}_{s,b,l}$ is the vector of genomic sequence context covariates, and $\boldsymbol{\beta}$ is the vector of fitted coefficients. The sequencing depth for strand *s* at location *l* is treated as the offset.

$$\log(\lambda_{s,b,l}) = \log\left(E(N_{s,b,l}|\boldsymbol{c}_{s,b,l})\right) = \log(d_{s,l}) + \boldsymbol{\beta}'\boldsymbol{c}_{s,b,l} \quad (8)$$

The negative binomial distribution GLM with log link function can be expressed in equation (9), where $\mu_{s,b,l}$ represents the expected mean for $N_{s,b,l}$ and θ is the dispersion parameter (the shape parameter of the gamma mixing distribution). The mean $E(N_{s,b,l}) = \mu_{s,b,l}$ and variance $VAR(N_{s,b,l}) = \mu_{s,b,l} + \theta\mu_{s,b,l}^{\ 2}$ can be estimated from GLM shown below.

$$\log(\mu_{s,b,l}) = \log\left(E(N_{s,b,l}|\boldsymbol{c}_{s,b,l})\right) = \log(d_{s,l}) + \boldsymbol{\beta}'\boldsymbol{c}_{s,b,l} \quad (9)$$

The zero-inflated Poisson distribution can be written as:

$$P\left(N_{s,b,l} = n_{s,b,l} | \pi_{s,b,l}, \lambda_{s,b,l}, \theta\right)$$

$$= \begin{cases} \pi_{s,b,l} + \left(1 - \pi_{s,b,l}\right) Pois\left(\lambda_{s,b,l}; 0\right) & if \;\; n_{s,b,l} = 0 \\ \left(1 - \pi_{s,b,l}\right) Pois\left(\lambda_{s,b,l}; n_{s,b,l}\right) & if \;\; n_{s,b,l} > 0 \end{cases} \tag{10}$$

Parameters of the zero-inflated Poisson distribution in equation (10) can be estimated by generalized linear model as shown in equation (11), where $\boldsymbol{z}_{s,b,l}$ is the vector of genomic sequence context covariates for the zero part, and $\boldsymbol{\gamma}$ is the vector of fitted coefficients.

$$\begin{gathered} \text{logit}\left(\frac{\pi_{s,b,l}}{1 - \pi_{s,b,l}}\right) = \boldsymbol{\gamma}' \boldsymbol{z}_{s,b,l} \\ \log\left(\lambda_{s,b,l}\right) = \boldsymbol{\beta}' \boldsymbol{c}_{s,b,l} \end{gathered} \tag{11}$$

The zero-inflated negative binomial distribution can be written as:

$$P\left(N_{s,b,l} = n_{s,b,l} | \boldsymbol{c}_{s,b,l}, \boldsymbol{z}_{s,b,l}\right)$$

$$= \begin{cases} \pi_{s,b,l} + \left(1 - \pi_{s,b,l}\right) NB\left(\mu_{s,b,l}, \theta; 0\right) & if \;\; n_{s,b,l} = 0 \\ \left(1 - \pi_{s,b,l}\right) NB\left(\mu_{s,b,l}, \theta; n_{s,b,l}\right) & if \;\; n_{s,b,l} > 0 \end{cases} \tag{12}$$

Parameters of the zero-inflated negative binomial distribution in equation (12) can be estimated by generalized linear model as shown in equation (13).

$$\begin{gathered} \text{logit}\left(\frac{\pi_{s,b,l}}{1 - \pi_{s,b,l}}\right) = \boldsymbol{\gamma}' \boldsymbol{z}_{s,b,l} \\ \log\left(\mu_{s,b,l}\right) = \boldsymbol{\beta}' \boldsymbol{c}_{s,b,l} \end{gathered} \tag{13}$$

4.2.4 Variant Identification

In chapter 2, the variant identification is done using Bayes factor by calculating the likelihood ratio of two models: M_E, the 'sequencing error distribution' model and M_V, the 'targeted lowest identifiable frequency distribution' model. Thus a predefined 'targeted lowest frequency' is needed. Here we used a hypothesis testing approach, to call candidate SNVs if the data are not from the sequencing error distribution. Specifically, a location with a certain alternative base is called as a candidate SNV if the numbers of reads from both strands are significantly greater than the predicted error rates. The p values were corrected using Benjamini–Hochberg procedure [146]. The corrected p value cut-off is 0.01.

4.2.5 Performance Evaluation Measurements

Precision and recall are defined as equations (4) and (5) in chapter 2. F1 score is defined below.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{14}$$

For Ion Proton dataset, same as chapter 2, only loci with at least 5 reads supporting alternative base are included in the evaluation.

4.3 Results

In the Result section, we first show the intuitions derived from visualization inspection for diagnosing the distribution form of sequencing error modeling. Then utilizing statistical testing of goodness-of-fit on different distributions, we selected the candidate distributions more appropriate for fitting the sequencing error count data. Then

we showed the performance of position specific error model using different distributions on identifying candidate SNVs, highlighted the merits of choosing appropriate distributions.

4.3.1 Candidate Statistical Distributions Selection

If the count data follow a given discrete distribution, then the visualization from distplot [141-143] shows the metameter is a linear function of all observed values. We plotted the number of reads containing non-reference alleles from all targeted region loci against binomial, Poisson and negative binomial distributions. As shown in Figure 10, the obvious curve for binomial distribution plot indicates the data do not follow binomial distribution. The plots for Poisson and negative binomial distributions show better agreement with the straight line although both curves deviate more from the straight line when the x-axis approaches 0. Further, if for each locus, the observed number of reads supporting each possible substitution type is called ‘error instance’. Then tabulating the percentages of zeros in all the error instances within the target regions, we got 85% from Ion Proton training dataset. Thus zero-inflated models should be considered. In the modeling step, we included Poisson, negative binomial and their zero-inflated counterparts (zero-inflated Poisson [147] and zero-inflated negative binomial [148]) as the candidate distributions under generalized linear model framework.

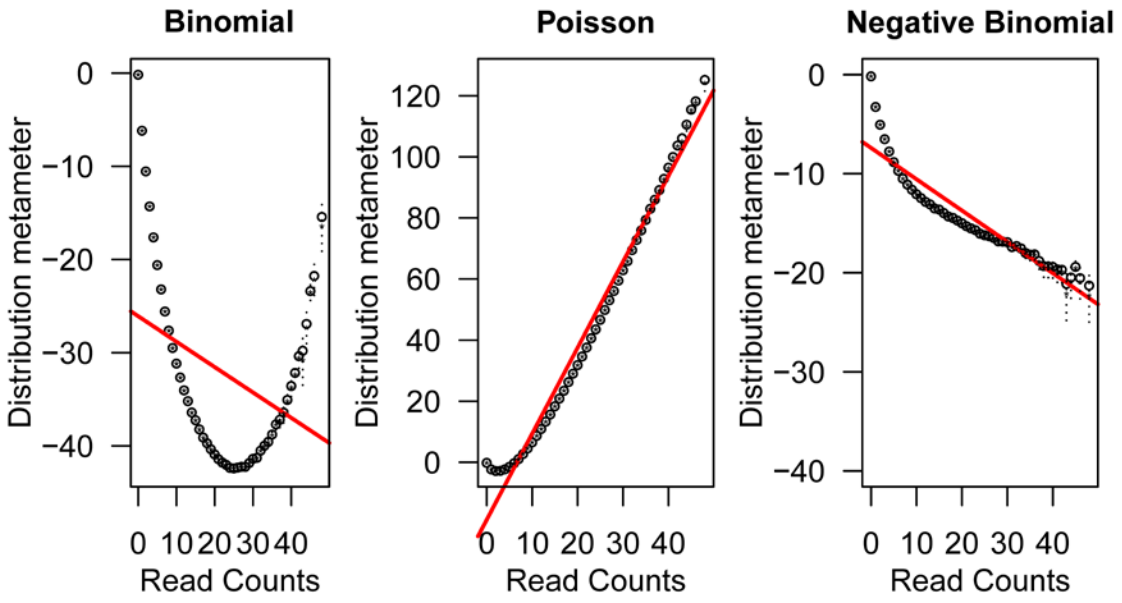

Figure 10 Distplot on binomial, Poisson and negative binomial distributions. The y-axis is the distribution metameter calculated by the method distplot used. The open points show the observed count metameters; the filled points show the confidence interval centers and the dashed lines show the confidence intervals for each point. 95% confidence interval is used.

4.3.2 Comparing the Goodness-of-Fit of Different Distributions

9 genomic sequence context covariates, totaling 24 degrees of freedom, were included in the generalized linear models (section 3.2.2 and Table 2). Since zero-inflated Poisson and zero-inflated negative binomial generalized linear models require covariates for both the 'zero' and 'count' parts, the same covariates were provided for both, resulting in doubled degrees of freedom of those included in Poisson and negative binomial generalized linear models.

To compare the goodness-of-fit of models based on different distributions, we used Vuong's non-nested hypothesis test [149]. BIC-corrected Vuong z-statistic [150] was used to impose stronger penalty on additional parameters. The pairwise comparison results are summarized in Table 11. Poisson distribution GLM is treated as the reference distribution to compare to, given its simple configuration. As expected, negative binomial GLM is superior to Poisson GLM, since negative binomial distribution models dispersion of the data, and this is also supported by dispersion test [151] ($z = 68.5881$, p value < 2.2e-16). The necessity of modeling zero-inflation is supported by the Vuong's test comparing zero-inflated Poisson with Poisson GLM. When comparing zero-inflated Poisson with negative binomial, negative binomial distribution fits the data better. However, it is worth noting the evidence of superiority – the absolute value of BIC-corrected Vuong z-statistic – is much smaller than the other tests. The merit of considering both dispersion and zero-inflation is further emphasized by the comparisons of zero-inflated Poisson with zero-inflated negative binomial and negative binomial with zero-inflated negative binomial. In conclusion, based on Vuong's test, for Ion Proton sequencing dataset, the most appropriate distribution for modeling DNA sequencing error read counts is zero-inflated negative binomial distribution.

Table 11 Vuong's non-nested tests for Ion Proton training data. NB: negative binomial. ZIP: zero-inflated Poisson. ZINB: zero-inflated negative binomial. '>' means a better fit of the left model.

Model 1	Model 2	Vuong z-statistic BIC-corrected	Hypothesis	P value
Poisson	NB	-122.67	model2 > model1	< 2.22e-16
Poisson	ZIP	-143.73	model2 > model1	< 2.22e-16
NB	ZIP	36.81	model1 > model2	< 2.22e-16
ZIP	ZINB	-92.16	model2 > model1	< 2.22e-16
NB	ZINB	-119.51	model2 > model1	< 2.22e-16

4.3.3 Performance Evaluation on Ion Proton Testing Benchmark

In the previous section, we used Vuong's test to establish that in modeling DNA sequencing error count data, the advantage of zero-inflated negative binomial distribution in fitting the data is statistically significant. In this section, we explored whether such statistical advantage could also be reflected in the ability to identify low-frequency SNVs. We first evaluated the overall precision and recall values of all distributions on the test benchmark. From Table 12, it is observed the Poisson GLM achieves the highest recall while zero-inflated negative binomial GLM has the highest precision. F1 score, the harmonic mean of precision and recall, is used to evaluate the overall performance. The conclusion from F1 score is consistent with that of Vuong's test, with zero-inflated negative binomial performs the best and is followed by negative binomial, zero-inflated Poisson and Poisson GLM. However, the precision values listed in Table 12 are lower than the ones reported previously [15, 72, 73]. There are 2 major reasons: 1. the Ion Proton test benchmark dataset is designed to enrich with low-frequency SNVs, with 68.9% of all SNVs of allele frequency <= 3%, in which 17.3% at 0.5% frequency and 19.8% at 1% frequency. Whereas the majority of previous studies focused on SNVs of >= 5% allele frequency; 2. one popular paradigm of SNV calling is a two-step procedure, first generating SNV candidates and then applying different methods to recalibrate the SNV call, for example filters and machine-learning based recalibration. The PSEM aims to efficiently recover high quality SNV candidates to facilitate the downstream candidate recalibration step, thus it is only fair to compare the performance of PSEM with other candidate generating methods.

Then we evaluated the effect of different variant identification methods. We compared the hypothesis-testing based variant identification with the Bayes factor approach used in chapter 2 (Poi_BF in Table 12). The overall F1 score for Bayes factor approach Poi_BF is between Poisson distribution and zero-inflated Poisson distribution,

and is notably inferior to negative binomial distribution and zero-inflated negative binomial distribution. In addition, hypothesis-testing approach does not require an additional parameter specifying the targeted lowest frequency required by Bayes factor approach. Thus with more appropriate distribution, not only higher performances but also a method with less additional constrains can be achieved. The result from VarScan2 before applying sequencing quality filters was included in Table 12. It is evident that except for Poisson GLM with hypothesis testing and Poisson GLM with Bayes factor, the other methods outperformed VarScan2 in both recall and precision. Therefore, choosing appropriate statistical modeling method enables us to recover more true SNVs without any loss of precision in candidate generating step.

Next, for all distributions, we explored the performance profiles on different allele frequencies. As shown in Figure 11, the clearly layered F1 score levels clearly show that SNVs of lower allele frequencies are more difficult to identify, no matter what distributions were used. In addition, the significant separation of 0.5% from the other allele frequencies indicate the detection limit is around 0.5% under current sequencing platform and depth. Meanwhile, the power of appropriate modeling is evident when comparing the performances of all distributions on SNVs of 0.5% allele frequency. Relative to Poisson GLM, considering either zero-inflation or dispersion boosted the F1 score by about 0.2 at 0.5%, while considering both by zero-inflated negative binomial further increased F1 score by about 0.1. Interestingly, compared with the second best model – negative binomial GLM, both precision and recall increased in zero-inflated negative binomial GLM, which pinpoints the necessity of modeling zero-inflation to derive more accurate error rates estimation. Furthermore, for SNVs with allele frequency greater than 1%, the average recall is 97.5% with 82.3% average precision for zero-inflated negative binomial GLM. Comparing the effect of different variant identification approaches, we can see although Poisson GLM performed better with Bayes factor than

with hypothesis testing, the differences in performances on less than 3% allele frequencies are evident compared with the most appropriate distribution zero-inflated negative binomial. To summarize, the performance evaluation results on low-frequency SNV identification also support the conclusion from Vuong's non-nested test, with zero-inflated negative binomial being the most appropriate model. Further, the necessity of modeling both dispersion and zero-inflation is exemplified by the much-elevated performance at close to sequencing error rate allele frequency, which is important for pushing down the detection limit of low-frequency SNV callers.

Table 12 Overall performance comparisons on Ion Proton testing benchmark. NB: negative binomial. ZIP: zero-inflated Poisson. ZINB: zero-inflated negative binomial. Poi_BF: GLM using Poisson distribution, Bayes factor approach used in chapter 2.

	Poisson	**NB**	**ZIP**	**ZINB**	**Poi_BF**	**VarScan2**
Recall	0.98	0.89	0.95	0.90	0.96	0.83
Precision	0.25	0.62	0.54	0.71	0.49	0.53
F1 Score	0.40	0.73	0.69	0.79	0.65	0.65

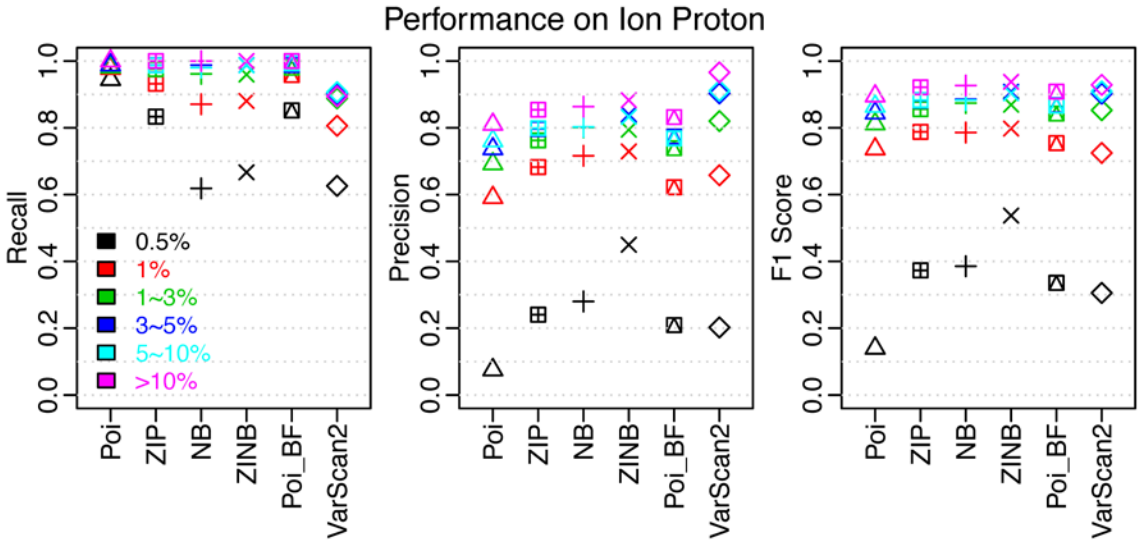

Figure 11 Performance by allele frequency Ion Proton testing benchmark. NB: negative binomial. ZIP: zero-inflated Poisson. ZINB: zero-inflated negative binomial. Poi_BF: position specific error model using Poisson distribution, with variant identification method being the Bayes factor approach used in chapter 2.

4.4 Discussion

The PSEM model aims to predict the position specific error rates associated with various genomic sequence contexts, under which the specific sequencing technology is prone to error. Based on publications evaluating features associated with sequencing errors and experiences from our previous effort, 9 types of significant features are considered. With the features fixed, using GLM, we evaluated the appropriateness of distributions with different mean – variance relationships and the ability to consider zero-inflation. Consistent with the computational tool EdgeR [152] for RNA-Seq data, we found the ability to model over-dispersion by negative binomial distribution necessary for DNA-Seq data as well. Additionally, for DNA-Seq error read counts modeling, zero-inflation is also a key factor for accurate prediction and inference. The much-elevated F1 score for 0.5% allele frequency SNVs as well as the highest overall performance by ZINB GLM highlighted the importance of choosing suitable statistical models. In addition, comparing different variant identification methods, we can see with the appropriate distribution, we can use a simple hypothesis-testing approach without requiring additional parameters required by Bayes factor, yet still can achieve a higher performance. Moreover, comparing with VarScan2, which conducts the Fisher's exact test for each targeted location on paired normal-tumor sequencing data, the significance of applying the correct reference error model is exemplified by higher recalls as well as precisions for 0.5% and 1% frequency SNVs. In theory, for low frequency SNV loci, VarScan2 treated the sequencing reads with non-reference bases from normal as the background error, which is essentially point estimation based on one location. Whereas PSEM collectively considers all loci with similar context features and thus is able to generate more accurate error estimation.

The current GLM-based PSEM framework only considers 9 types of genome sequence context features. To further improve the performances, more informative

features associated with sequencing errors should be included and tested. In addition, from the modeling aspect, exploration of the potential to further increase the performances by applying more sophisticated computational models are desired. When applying the position specific error model on sequencing platforms different from Ion Proton, new features related to the underlying biochemistry might be added. Thus, to better understand its generalizability and adaptiveness in the features used, tests on other sequencing technologies, such as Illumina, SOLiD and Complete Genomics, are necessary. Except for the sequencing platform effect, the effects of other steps in the sequencing library preparation should also be considered, for example the target capture assay. Since the capture assay for the Ion Proton benchmarks is amplicon-based, thus the reads from the same amplicon are supposed to have the same start and end locations. However, hybridization-based approach tends to generate reads with different start and end location, therefore, it should be tested to compare the performance profiles with amplicon-based approached to see if such a difference may impact the position specific error modeling.

Differentiating low frequency SNVs from sequencing artifacts is the key for identifying SNVs at frequencies close to sequencing error rates. Our PSEM approach tried to push the limit toward the sequencing error rates. Based on the analyses on benchmarks from standard sequencing protocols and the given sequencing depth, we speculate the detection limit is around 0.5% on the regions covering all exons of hundred of genes, with a total size up to millions of bases. However, with high accuracy sequencing protocols, such as duplex sequencing [64] and ultra-deep target enrichment assay [97], the researchers reported identification of SNVs around 0.1% on a single gene scale. For the future direction, it is worthwhile to test whether we can push the detection limit below 0.5% or even 0.1% by coupling the improved experimental protocols with our position specific error modeling.

Chapter 5. Statistical Modeling for Illumina MiSeq Platform Genomic Sequence Context Dependent Error

5.1 Introduction

In chapter 4, we explored the possibility of improving the performances of position specific error modeling with different statistical distributions. However, same as chapter 2, all tests were done on Ion Proton sequencing platform. To understand how position specific error modeling behaves on different sequencing platforms, we replicated the analysis conducted in chapter 4 on a publically available Illumina MiSeq benchmark dataset, published with the low-frequency SNV detection method UDT-Seq [15]. The Illumina MiSeq benchmark dataset was chosen for several considerations. First, different from the semiconductor based sequencing utilized by Ion Proton sequencers; Illumina sequencing platforms used optical system based sequencing by synthesis (SBS). This difference enables us to check whether generalized linear model based position specific error model can adapt to sequencing platforms based on completely different biochemistries. Second, similar to the Ion Proton benchmarks, the amplicon based target capture assay was also used by the Illumina MiSeq benchmark dataset. Thus we have one less major complication in interpreting the differences. Third, the lowest targeted frequency for Illumina MiSeq benchmark is 1%, thus allowing us to characterize the effect of different distributions on low-frequency range.

With the Illumina MiSeq benchmark data set, we wanted to examine whether the generalized linear model can be utilized and if so, whether there are any differences in terms of the contribution of different sequence context features. For this purpose, we controlled the sequence context features to be the same as the ones used in Ion Proton benchmarks. In addition, whether the most appropriate statistical distribution remains to be zero-inflated negative binomial distribution.

5.2 Materials and Methods

5.2.1 Illumina MiSeq Benchmark Dataset Overview

The design details can be found in the UDT-Seq paper [15]. Briefly, the length of targeted regions for Illumina MiSeq datasets is 23.2 kb, covered by 158 amplicons with about 200-nucleotide long. The amplification was done with microdroplet PCR [153]. This Illumina MiSeq benchmark data were generated by mixing 4 individuals at 4 different percentages and then permuted the mixing percentage assignment 4 times to generate 4 calibration datasets – CAL_A, CAL_B, CAL_C and CAL_D, details shown in Table 13. Sequencing was done with Illumina MiSeq platform, and the read is 151-nucleotide long. The raw reads were downloaded from NCBI Short Read Archive [154, 155] (SRP009487.1) and processed as the paper described. Reads with mapping quality less than 30 were filtered out.

For the choice of Illumina MiSeq training and testing benchmarks, since the 4 calibration data sets were generated with the same procedures, without loss of generality, we used CAL_A as training benchmark and treated the others as testing benchmark. Also, different from Ion Proton benchmarks, these benchmarks are similar to the tumor only or pooled sequencing samples. Thus the all identified candidate SNVs were used in the performance evaluation.

5.2.2 Generalized Linear Models and Variant Identification

To test how different sequencing platforms impact the generalized linear model fitting, the same sequence context features used in chapter 4 were also used on Illumina MiSeq data. The details of the 9 genomic sequence contexts considered in generalized linear models were summarized in Table 2. The same 4 statistical distributions were fitted: Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial. In addition, the same hypothesis testing based variant identification testing in

chapter 4 was also used here, requiring Benjamini–Hochberg procedure corrected p values from both strands to be less than 0.01.

Table 13 Illumina MiSeq benchmark design.

ID	CAL_A	CAL_B	CAL_C	CAL_D
NA12156	1%	5%	20%	74%
NA12878	5%	20%	74%	1%
NA18507	20%	74%	1%	5%
NA19240	74%	1%	5%	20%

5.2.3 Performance Evaluation Measurements

Precision, recall and F1 score were used for performance evaluation. For Illumina MiSeq dataset, filter 2 used by UDT-Seq [15] was applied which requires >= 0.2% frequency for alternative bases. However, the other filters were not used, including filter 1 removing positions within primers, filter 3 position in the read, filter 4 depth strand bias, filter 5 depth discrepancies between training and testing samples, filter 6 binomial test p values on the significance of different from sequencing error rates and filter 7 local sequencing context based filters. We relied on the PSEM framework to properly address sequence contexts and depth related problems.

5.3 Results

In the result section, we first show the comparison of the appropriateness of the 4 candidate distributions. Then we compared the coefficients in generalized linear models on Illumina MiSeq with those from Ion Proton to look for the impact of different sequencing benchmarks. Next, to set the stage for understanding the performance differences, we first compared the differences in designed allele frequency composition as well as sequencing depth between the Illumina MiSeq and Ion Proton testing benchmarks. We concluded by evaluating the candidate variant identification performances and the comparison between position specific error model and UDT-Seq.

5.3.1 Comparing the Goodness-of-Fit of Different Distributions

To evaluate the generalizability and adaptiveness of the generalized linear model based position specific error modeling, the same modeling strategies were applied to the Illumina MiSeq sequencing data sets. Similar to the analysis on Ion Proton data set, paired Vuong's non-nested hypothesis tests were conducted on the 4 candidate distributions, with details summarized in Table 14. The tests show the most appropriate

distribution is still zero-inflated negative binomial. However, for the negative binomial (NB) (model 1) and zero-inflated Poisson (ZIP) (model 2) comparison, the BIC-corrected Vuong z-statistic is -0.47 resulting in p value = 0.318. Therefore the goodness-of-fit for these two distributions on MiSeq dataset are not significantly different.

5.3.2 Comparing Generalized Linear Models on Different Sequencing Platforms

Despite similar statistical modeling schema can be readily generalized to Illumina MiSeq data set, Illumina MiSeq and Ion Proton sequencers differ significantly in terms of sequencing chemistry. The former is based on sequencing-by-synthesis (SBS) that relies on high-resolution optic systems, whereas the latter is based on Ion semiconductor sequencing where no modified nucleotides or optics are required. The differences in sequencing mechanisms make Ion Proton sequencers run faster but are prone to homopolymer related errors. Comparing the negative binomial generalized linear model regression coefficients on both datasets (Table 15), homopolymer related features significant in Ion Proton data set regression are either insignificant (hmer_len, hmer_dist) or show opposite effect (hmer_op, hmer_den) on the error rate.

5.3.3 Benchmarks Comparison

Comparing the Illumina MiSeq testing benchmark with the Ion Proton testing benchmark, Ion Proton dataset contains a total of 1557 somatic SNVs while Illumina MiSeq dataset contains 514 SNVs, in which 175 SNVs are unique. More importantly, Ion Proton benchmark was designed to comprehensively characterize the SNV caller performance on close to sequencing error allele frequencies, thus it is enriched with SNVs of <= 3% allele frequencies, with 0.5% as the lowest targeted frequency. Plotting the cumulative percentages of SNV numbers at different allele frequencies (Figure 12) from the two test benchmarks, it is clear the major components of Ion Proton test

benchmark SNV allele frequencies are at 0.5%, 1%, 2% to 5%, followed by continuous frequencies until 46%, the maximum somatic SNV frequency designed in the dataset. Whereas Illumina MiSeq testing benchmark set is enriched with SNVs at the 4 discrete allele frequency levels same as the design.

Table 14 Vuong's non-nested test on 4 distributions applied to Illumina MiSeq training data.

Model 1	Model 2	Vuong z-statistic BIC-corrected	Hypothesis	P value
Poisson	NB	-23.38	model2 > model1	< 2.22e-16
Poisson	ZIP	-21.30	model2 > model1	< 2.22e-16
NB	ZIP	-0.47	model2 < model1	0.31796
ZIP	ZINB	-20.22	model2 > model1	< 2.22e-16
NB	ZINB	-17.44	model2 > model1	< 2.22e-16

Table 15 Negative binomial GLM coefficients for Ion Proton and Illumina MiSeq training datasets.

	Ion Proton			Illumina MiSeq		
Parameter	Estimate	Standard Error	P value	Estimate	Standard Error	P value
(Intercept)	-10.9331	0.0105	< 2e-16	-1.0700	0.0290	< 2e-16
A → C	-0.1516	0.0072	< 2e-16	0.6922	0.0220	< 2e-16
A → G	1.5093	0.0060	< 2e-16	2.6550	0.0190	< 2e-16
A → T	-0.0352	0.0072	< 2e-16	-0.2219	0.0270	4.26e-16
C → A	-0.1073	0.0075	< 2e-16	-0.3068	0.0290	< 2e-16
C → G	-0.3362	0.0080	< 2e-16	-0.9909	0.0360	< 2e-16
C → T	1.3287	0.0062	< 2e-16	1.8840	0.0200	< 2e-16
G → A	1.2600	0.0062	< 2e-16	1.7870	0.0200	< 2e-16
G → C	-0.2030	0.0077	< 2e-16	-0.7178	0.0330	< 2e-16
G → T	-0.1705	0.0075	< 2e-16	-0.7540	0.0360	< 2e-16
T → A	0.0445	0.0071	< 2e-16	-0.2594	0.0270	< 2e-16
T → C	1.5362	0.0059	< 2e-16	2.7120	0.0190	< 2e-16
up base A	0.1046	0.0037	< 2e-16	0.1640	0.0100	< 2e-16
up base C	0.2316	0.0038	< 2e-16	0.4087	0.0100	< 2e-16
up base G	0.3288	0.0037	< 2e-16	0.5110	0.0100	< 2e-16
down base A	-0.0908	0.0037	< 2e-16	-0.2120	0.0100	< 2e-16
down base C	0.3356	0.0036	< 2e-16	0.3462	0.0090	< 2e-16
down base G	0.1748	0.0037	< 2e-16	0.2600	0.0090	< 2e-16
GC	0.0058	0.0001	< 2e-16	0.0099	0.0003	< 2e-16
hmer_den	0.2994	0.0136	< 2e-16	-0.0798	0.0360	0.028
hmer_op	0.3313	0.0037	< 2e-16	-0.1430	0.0100	< 2e-16
hmer_dist	-0.0137	0.0004	< 2e-16	-0.0009	0.0010	0.4
hmer_len	0.0790	0.0012	< 2e-16	0.0018	0.0030	0.569

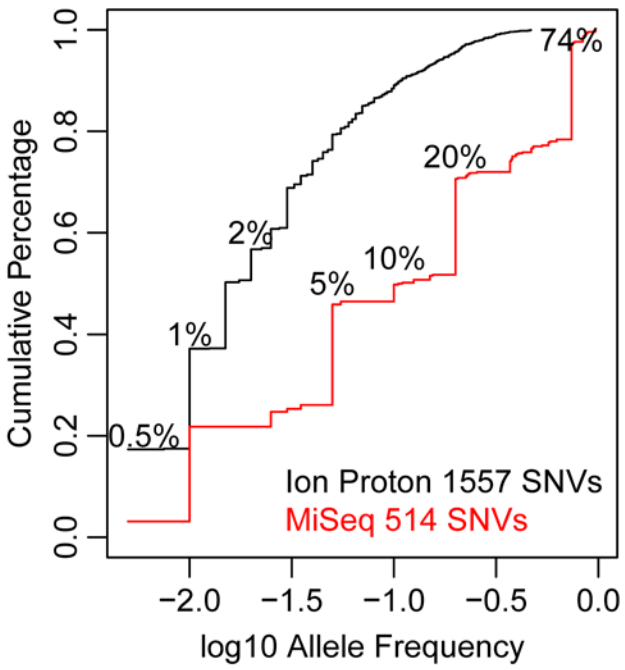

Figure 12 Allele frequency composition of Ion Proton and Illumina MiSeq testing benchmark SNVs.

Except for allele frequency composition, sequencing depth is also a crucial factor affecting the performances of the SNV callers, especially at the low-frequency range. The average depth for Ion Proton sequencing testing benchmark is about 4000x and about 1500x for MiSeq. In addition, despite the amplicon-based capture assay was applied on benchmark datasets from both technologies, the evenness of the depth across the targeted regions is different. When comparing the depth on known testing benchmark SNV loci of two technologies (Figure 13), the depth distribution for Ion Proton is skewed while the distribution profile for Illumina MiSeq data displays a bell shape. For Ion Proton testing benchmark SNV loci, 85.4% of all loci have a depth no less than 1000x, while 98.1% for Illumina MiSeq. Further, the average depth at SNV loci from both benchmarks are around 3000x, despite the much higher overall depth in Ion Proton benchmark. Thus, we speculate lowered recall for some Ion Proton benchmark SNVs, particularly for the $\leq$ 1% ones, the identifiable power of which are more sensitive to the depth and read count number sampling variances.

5.3.4 Performance Evaluation on Illumina MiSeq Testing Benchmark

To evaluate whether the differences in generalized linear model coefficients affect the performance profiles on various allele frequencies, we applied the 4 generalized linear models trained on CAL_A to the testing benchmark dataset combining CAL_B, CAL_C and CAL_D. And then we conducted the recall, precision and F1 score analyses by allele frequency on the combined dataset. As shown in Figure 14, similar to the Ion Proton data set, SNVs of lower allele frequencies are more difficult to identify. However, when comparing the performances of zero-inflated Poisson with negative binomial GLM on 0.5% ~ 1% allele frequency, different from Ion Proton dataset, negative binomial demonstrated a much higher F1 score compared with zero-inflated Poisson. A closer look at the performance profiles shows the noticeable drop in recall comparing

negative binomial with zero-inflated Poisson in Ion Proton is absent in MiSeq data. Examination on the benchmark SNVs missed by negative binomial but recovered by zero-inflated Poisson showed lower depth for the missed ones. Therefore the absent of recall drop in MiSeq is due to its relatively even depth contrast to the Ion Proton dataset (Figure 13). For SNVs with > 1% allele frequency, the F1 scores are all greater than 0.9 and clustered together for all distributions.

Comparing with the results from UDT-Seq [15], which reported approximately 90% recall and >95% precision (no specific number was given, the precision was inferred by the precision for the other data UDT-Seq tested - Illumina GAII benchmark data at 1500x depth), zero-inflated negative binomial generalized linear model demonstrates higher overall recall (95.1%) and high precision (93.4%).

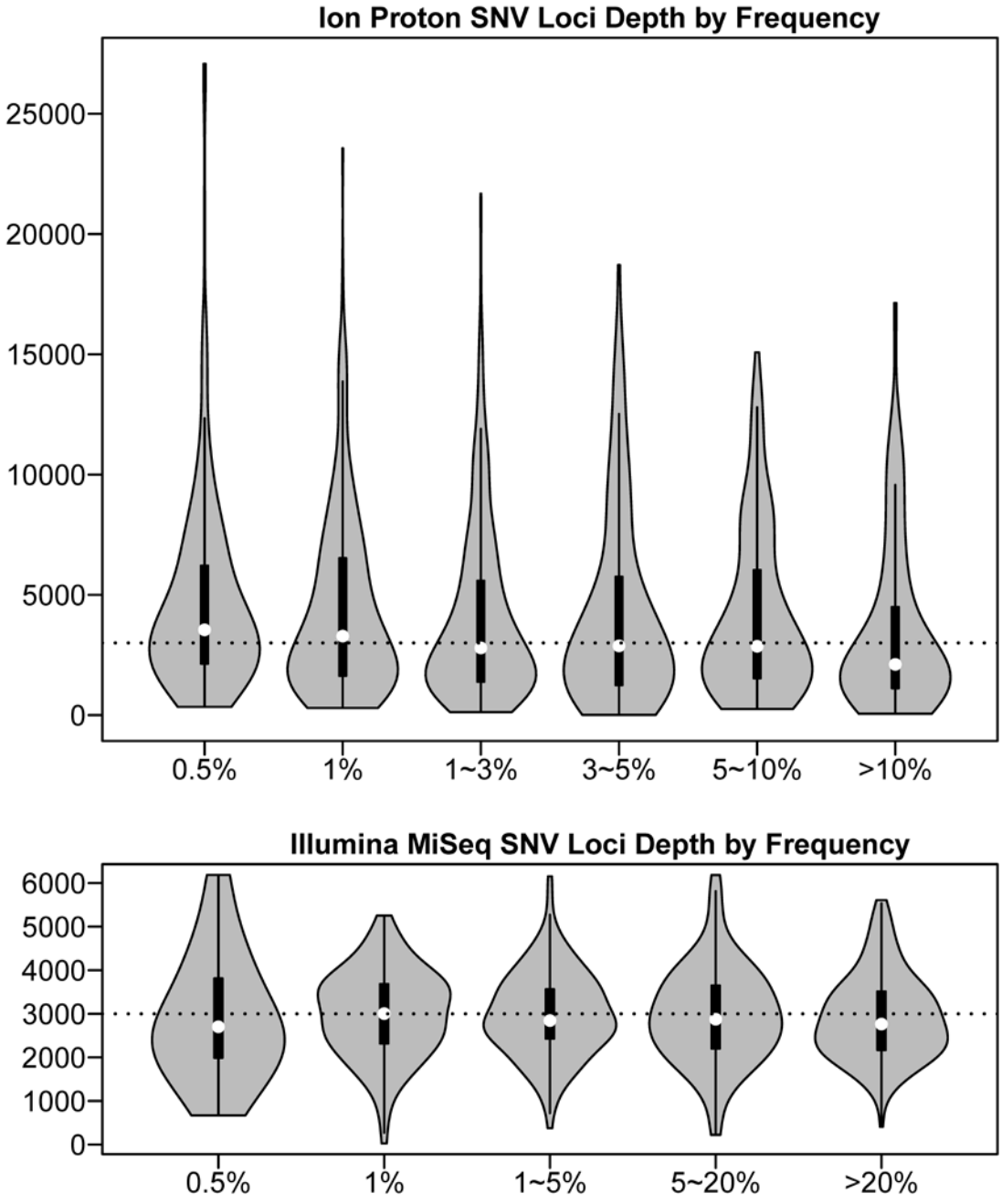

Figure 13 SNV loci depth distribution by allele frequency for Ion Proton and Illumina MiSeq. The dashed lines show the 3000x depth.

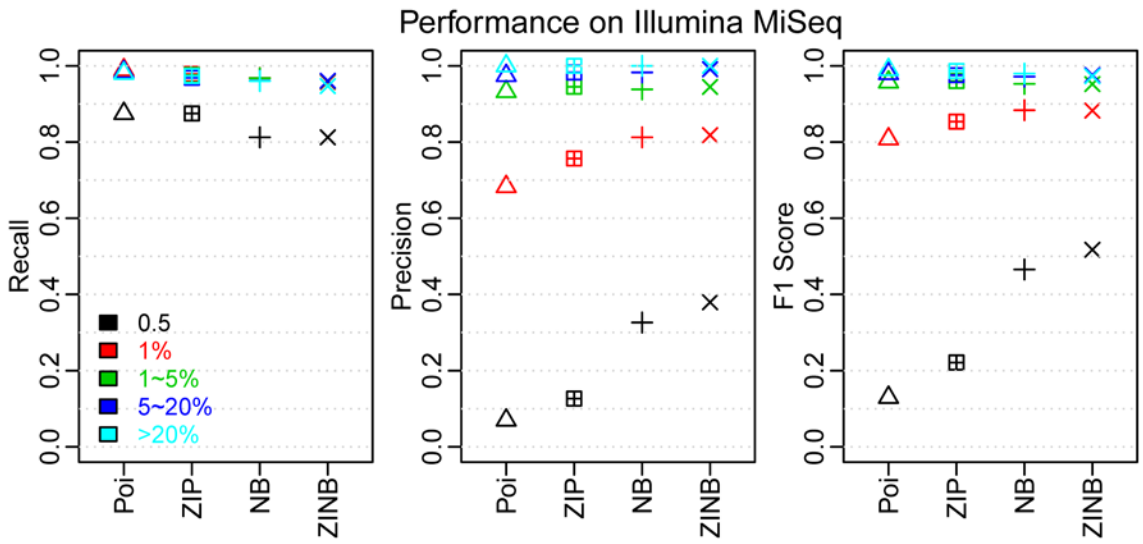

Figure 14 Performance by allele frequency summary on Illumina MiSeq testing benchmark.

5.4 Discussion

The evaluation of position specific error modeling on Illumina MiSeq dataset showed the generalizability of the position specific error modeling framework as well as its adaptiveness to different technologies. The position specific error modeling framework adapts to training data from different technologies by adjusting the coefficients in fitted generalized linear models. Moreover, except for the established importance of choosing appropriate statistical model, the sequencing depth evenness is also an important factor affecting low-frequency SNVs calling performances.

Chapter 6. Conclusions and Future Directions

The overall focus of this dissertation project is to develop a computational framework to sensitively and specifically detect low frequency SNVs from NGS based DNA sequencing data. The major difficulty of this task is to characterize artifacts in sequencing data and distinguish them from the low-frequency SNVs, which may present at a similar level as the artifacts. We sought to tackle this problem by modeling the end result of all sources of errors originated from various steps of the NGS experiment workflow, to effectively distinguish errors from low-frequency SNVs computationally. Particularly, the position specific error model characterizes the genomic sequence contexts-dependent error tendencies of the sequencers and thus determines the detection limits for sensitive low-frequency SNV identification. Machine-learning-based recalibration further considers sequencing quality features unique to each candidate SNV locus and boosts the specificity. Training data containing comprehensive low-frequency SNVs are needed for the computational framework to build a representative and robust model. Since different sequencing pipelines (generally including library preparation, amplification and sequencing instrument) have differential error profiles [156], it is a challenging task to develop a model that adapts to different sequencing pipelines. The configuration of RareVar computational framework enables it to adjust to different pipelines when fed with data from the target pipeline. This benchmark generation step is an integral part of RareVar framework when applied to data from a previously uncharacterized sequencing pipeline. By performance comparison with existing methods, we confirmed the effectiveness of RareVar in sensitively and reliably detecting low-frequency SNVs, with the advantage most evident in 0.5% ~ 3% allele frequencies.

Generating a benchmark dataset suitable for comprehensive low-frequency SNVs detection evaluation is a nontrivial task [82]. By sequencing the DNA sample mixture of multiple individuals previously genotyped by 1000 Genomes Project, the benchmark data not only contain a large number of known SNVs, but also preserve the bona fide sequencer error profiles likely lost by simulation-based approaches [76-80, 82]. Also, contrast to other deep sequencing efforts for low-frequency SNV detection, which targeted less than 50kb regions, Ion AmpliSeq Comprehensive cancer panel targets about 1.7 million bases, encompassing half of the known oncogenes and tumor suppressor genes. The invariant loci of the targeted regions provide comprehensive training data for the position specific error model, which facilitate the building of unbiased error profile models. On the other hand, the large number of variant loci, which were enriched with low-frequency SNVs by our design, allows the machine-learning-based recalibration to delineate the sophisticated boundaries between true low-frequency SNVs and sequencing artifacts.

The ability to capture the differential context-dependent error rates is the key for sensitive detection of close to error rate SNVs. We applied Poisson-distributed generalized linear model to integrate 9 sequence context features for position specific error rates modeling. Comparing with existing error rate modeling approaches, our generalized linear model framework modeled the combinatorial effects of more features than tabulation-based method [15], likelihood ratio based method [23, 72] and recursive sequencing error probability modeling [75], in which the latter two methods mainly rely on base quality feature. The position specific error model allows finer differentiation of biased sequencing error rates. Besides, the scalability of the generalized linear model removes the need to make unrealistic assumptions, such as the equal substitution error rates at each locus assumed by the likelihood ratio method. In the precision and recall comparisons with other tools, the position specific error model recovered the most

known SNVs, with the advantage more evident at lower allele frequency ranges. In addition, its precision is comparable with VarScan2, thus selecting a high quality candidate SNV set for further refinement.

The machine-learning-based candidate recalibration step in RareVar considers sequencing quality features to refine the candidates. This strategy was also used in GATK [65]. However, the recalibration in our framework is tailored to low-frequency SNV detection. First, comparing to the dbSNP germline SNVs used as positive set in GATK, the true SNVs in our designed benchmark covered a wide range of continuous frequencies and more importantly, enriched at low frequency ranges. Thus it is a suitable training data set for modeling features of cancer somatic mutations and pooled sequencing variants, especially at the lower frequencies. Second, instead of trying to only capture the characteristics of true SNVs in GATK, we had both true SNVs defined by the benchmark data as well as the false positive SNVs generated by position specific error model step, allowing us to distinguish the two types. Moreover, the candidate SNVs derived after position specific error modeling are enriched with true SNVs, thus they constitute an ideal training set to optimize the classification boundaries for higher sensitivity and specificities. We showed the effectiveness of the machine-learning-based recalibration in boosting the precision as well as preserving high recall by comparing it with the position specific error model as well as other existing tools. The aforementioned advantages in framework design were highlighted by the highest precision increase at 0.5% and 1% frequency ranges.

From benchmark design, position specific error model to machine-learning-based candidate recalibration, these major components of RareVar framework operate synergistically to optimize the performance on low-frequency SNV detection. Enriched low-frequency SNV benchmark enables supervised learning for the downstream components to effectively distinguish low-frequency SNVs from sequencing artifacts.

The strategy of using benchmark data also enables adapting to different sequencing platforms by feeding downstream components with training data from the platform of interest. The value of tailoring the model toward the platform used is pinpointed when RareVar and TVC, the two methods more tailored to Ion Torrent sequencing technology, significantly outperformed popular methods previously tailored for Illumina sequencing technology. Computational components position specific error model and machine-learning-based candidate recalibration characterize the context-dependent systematic sequencer error tendency and locus-specific sequencing qualities, respectively. Moreover, both the generalized linear model and machine learning algorithm random forest are capable of incorporating more features, thus guarantees extensibility of RareVar.

Next-generation sequencing error data are in essence count data. In chapter 3, we showed that the effectiveness of different statistical distributions on position specific error modeling was different. By keeping the sequence context features the same, observed differences in performance were due to differential goodness-of-fit for the tested distributions. Similar to RNA sequencing differential expression analysis, negative binomial distribution showed statistically significant better goodness-of-fit than Poisson distribution, due to its extra parameter in dealing with overdispersion of next-generation sequencing count data [152, 157]. Unique to sequencing error count data is the large percentage of zeros, or zero-inflation, since the average error rate is only 0.1% to 1% for most platforms. The zero-inflated counterparts of both Poisson and negative binomial distributions fit statistically better, as shown by Vuong's test. Zero-inflated negative binomial distribution statistically fit the sequencing error data the best. Such an advantage was also reflected in higher overall performance in detecting SNVs, especially at 0.5% and 1% frequencies, demonstrating the practical value of applying statistically fitter distributions. Furthermore, both Ion Proton and Illumina MiSeq data

supported this conclusion. Zero-inflated negative binomial distribution captures generic features of the next-generation sequencing error count data.

The adaptiveness of generalized linear model for position specific error modeling was demonstrated in chapter3. When fed with sequencing data from different platforms, the model adjusted to different platforms by learning different coefficients. Comparing the fitted coefficients for Ion Proton and Illumina MiSeq data, most homopolymer related features that explain Ion Proton sequencing errors were no longer significant in Illumina MiSeq model. Given the successful application of machine-learning-based recalibration in GATK, which is broadly applied on Illumina sequencing data, as well as its effectiveness in Ion Proton dataset demonstrated in chapter 2, both computational components in RareVar are proved to be adaptive to different sequencing platforms.

Low-frequency SNVs detection is the key component in identifying mutational drift and/or enrichment in breast tumors. By comparing the SNVs in primary tumor tissue with the ones identified in cultured reprogrammed tumor cells, low-prevalent and potentially actionable SNVs missed by sequencing bulk tumor could be recovered. Applying RareVar on these data, we identified both known and novel somatic mutations enriched in reprogrammed tumor cells. Further experimental validation and functional study is on going.

The future directions for RareVar framework refinement are improving the position specific error model and the machine-learning-based recalibration. For the position specific error model, identifying and incorporating more features informative of sequencer error tendencies are desired. Also, the benefit of considering interactions between features in addition to combinatorial effects is worth exploring. Similarly, designing more sequencing quality features as well as selecting the most informative subset of features for distinguishing true SNVs from sequencing artifacts are the major concerns for the machine-learning-based recalibration.

In summary, we developed an adaptive and flexible framework for high performance low-frequency SNV detection. Such a method extends the application territory of sequencing based strategies, and also have the potential to greatly facilitate cancer and population genetics researches as well as clinical applications such as cancer early diagnosis, metastasis monitoring and relapse identification.

REFERENCES

1. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB et al: An integrated map of genetic variation from 1,092 human genomes. Nature 2012, 491(7422):56-65.
2. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding MJ, Bamford S, Cole C, Ward S et al: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015, 43(D1):D805-D811.
3. Ventola CL: Role of pharmacogenomic biomarkers in predicting and improving drug response: part 1: the clinical significance of pharmacogenetic variants. P & T : a peer-reviewed journal for formulary management 2013, 38(9):545-560.
4. Iau PT, Macmillan RD, Blamey RW: Germ line mutations associated with breast cancer susceptibility. European journal of cancer 2001, 37(3):300-321.
5. Goldstein AM: Familial melanoma, pancreatic cancer and germline CDKN2A mutations. Human mutation 2004, 23(6):630.
6. Rieder H, Bartsch DK: Familial pancreatic cancer. Familial cancer 2004, 3(1):69-74.
7. Rogers CD, van der Heijden MS, Brune K, Yeo CJ, Hruban RH, Kern SE, Goggins M: The genetics of FANCC and FANCG in familial pancreatic cancer. Cancer biology & therapy 2004, 3(2):167-169.
8. Meindl A, Hellebrand H, Wiek C, Erven V, Wappenschmidt B, Niederacher D, Freund M, Lichtner P, Hartmann L, Schaal H et al: Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. Nat Genet 2010, 42(5):410-414.
9. Grant RC, Selander I, Connor AA, Selvarajah S, Borgida A, Briollais L, Petersen GM, Lerner-Ellis J, Holter S, Gallinger S: Prevalence of germline mutations in cancer predisposition genes in patients with pancreatic cancer. Gastroenterology 2015, 148(3):556-564.
10. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA et al: Germline Mutations in Predisposition Genes in Pediatric Cancer. The New England journal of medicine 2015.
11. Oxnard GR, Nguyen KS, Costa DB: Germline mutations in driver oncogenes and inherited lung cancer risk independent of smoking history. Journal of the National Cancer Institute 2014, 106(1):djt361.
12. Gaudet MM, Kirchhoff T, Green T, Vijai J, Korn JM, Guiducci C, Segre AV, McGee K, McGuffog L, Kartsonaki C et al: Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. PLoS genetics 2010, 6(10):e1001183.
13. Yin J, Yin M, Vogel U, Wu Y, Yao T, Cheng Y, Sun Z, Hou W, Wang C: NFKB1 common variants and PPP1R13L and CD3EAP in relation to lung cancer risk in a Chinese population. Gene 2015, 567(1):31-35.
14. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI: Evolutionary evidence of the effect of rare variants on disease etiology. Clinical genetics 2011, 79(3):199-206.
15. Harismendy O, Schwab RB, Bao L, Olson J, Rozenzhak S, Kotsopoulos SK, Pond S, Crain B, Chee MS, Messer K et al: Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome biology 2011, 12(12):R124.
16. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B: Detection and quantification of rare mutations with massively parallel sequencing. Proceedings

of the National Academy of Sciences of the United States of America 2011, 108(23):9530-9535.
17. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang ZM, Chanock SJ, Fraumeni JF, Chatterjee N: Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proceedings of the National Academy of Sciences of the United States of America 2011, 108(44):18026-18031.
18. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA: Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences of the United States of America 2012, 109(36):14508-14513.
19. Crowley E, Di Nicolantonio F, Loupakis F, Bardelli A: Liquid biopsy: monitoring cancer-genetics in the blood. Nat Rev Clin Oncol 2013, 10(8):472-484.
20. Stratton MR, Campbell PJ, Futreal PA: The cancer genome. Nature 2009, 458(7239):719-724.
21. Stratton MR: Exploring the Genomes of Cancer Cells: Progress and Promise. Science 2011, 331(6024):1553-1558.
22. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA et al: Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012, 30(5):413-+.
23. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013, 31(3):213-219.
24. Meacham CE, Morrison SJ: Tumour heterogeneity and cancer cell plasticity. Nature 2013, 501(7467):328-337.
25. Burrell RA, McGranahan N, Bartek J, Swanton C: The causes and consequences of genetic heterogeneity in cancer evolution. Nature 2013, 501(7467):338-345.
26. Marchetti A, Felicioni L, Buttitta F: Assessing EGFR mutations. New Engl J Med 2006, 354(5):526-527.
27. Ruiz MIG, Floor K, Rijmen F, Grunberg K, Rodriguez JA, Giaccone G: EGFR and K-ras mutation analysis in non-small cell lung cancer: Comparison of paraffin embedded versus frozen specimens. Cell Oncol 2007, 29(3):257-264.
28. Crowley E, Di Nicolantonio F, Loupakis F, Bardelli A: Liquid biopsy: monitoring cancer-genetics in the blood. Nature reviews Clinical oncology 2013, 10(8):472-484.
29. Bettegowda C, Sausen M, Leary R, Kinde I, Agrawal N, Bartlett B, Wang H, Luber B, Kinzler K, Vogelstein B et al: Detection of circulating tumor DNA in early and late stage human malignancies. Cancer Res 2014, 74(19).
30. Mego M, De Giorgi U, Dawood S, Wang X, Valero V, Andreopoulou E, Handy B, Ueno NT, Reuben JM, Cristofanilli M: Characterization of metastatic breast cancer patients with nondetectable circulating tumor cells. International journal of cancer Journal international du cancer 2011, 129(2):417-423.
31. Pierga JY, Hajage D, Bachelot T, Delaloge S, Brain E, Campone M, Dieras V, Rolland E, Mignot L, Mathiot C et al: High independent prognostic and predictive value of circulating tumor cells compared with serum tumor markers in a large prospective trial in first-line chemotherapy for metastatic breast cancer patients. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 2012, 23(3):618-624.
32. Roschewski M, Dunleavy K, Pittaluga S, Moorhead M, Pepin F, Kong K, Shovlin M, Jaffe ES, Staudt LM, Lai C et al: Circulating tumour DNA and CT monitoring in

patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. The Lancet Oncology 2015, 16(5):541-549.
33. Schwarzenbach H, Hoon DSB, Pantel K: Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer 2011, 11(6):426-437.
34. Diehl F, Li M, Dressman D, He Y, Shen D, Szabo S, Diaz LA, Jr., Goodman SN, David KA, Juhl H et al: Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(45):16368-16373.
35. SNP
36. Sham P, Bader JS, Craig I, O'Donovan M, Owen M: DNA pooling: A tool for large-scale association studies. Nat Rev Genet 2002, 3(11):862-871.
37. Bansal V, Tewhey R, Leproust EM, Schork NJ: Efficient and cost effective population resequencing by pooling and in-solution hybridization. PloS one 2011, 6(3):e18353.
38. Golan D, Erlich Y, Rosset S: Weighted pooling-practical and cost-effective techniques for pooled high-throughput sequencing. Bioinformatics 2012, 28(12):I197-I206.
39. Futschik A, Schlotterer C: The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. Genetics 2010, 186(1):207-218.
40. Gautier M, Foucaud J, Gharbi K, Cezard T, Galan M, Loiseau A, Thomson M, Pudlo P, Kerdelhue C, Estoup A: Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol Ecol 2013, 22(14):3766-3779.
41. Schlotterer C, Tobler R, Kofler R, Nolte V: Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nat Rev Genet 2014, 15(11):749-763.
42. Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010, 11(6):415-425.
43. Saint Pierre A, Genin E: How important are rare variants in common disease? Brief Funct Genomics 2014, 13(5):353-361.
44. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al: Finding the missing heritability of complex diseases. Nature 2009, 461(7265):747-753.
45. An Introduction to Next-Generation Sequencing Technology [http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf]
46. Ion Torrent™ Next-Generation Sequencing Technology [https://http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html]
47. Lin H, Zhang Z, Zhang MQ, Ma B, Li M: ZOOM! Zillions of oligos mapped. Bioinformatics 2008, 24(21):2431-2437.
48. Li H, Ruan J, Durbin R: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008, 18(11):1851-1858.
49. Jiang H, Wong WH: SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 2008, 24(20):2395-2396.
50. Homer N, Merriman B, Nelson SF: BFAST: an alignment tool for large scale genome resequencing. PloS one 2009, 4(11):e7767.
51. Li R, Li Y, Kristiansen K, Wang J: SOAP: short oligonucleotide alignment program. Bioinformatics 2008, 24(5):713-714.

52. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.
53. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics 2012, 13:341.
54. Robasky K, Lewis NE, Church GM: The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 2014, 15(1):56-62.
55. Yost SE, Smith EN, Schwab RB, Bao L, Jung H, Wang X, Voest E, Pierce JP, Messer K, Parker BA et al: Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. Nucleic Acids Res 2012, 40(14):e107.
56. Akbari M, Hansen MD, Halgunset J, Skorpen F, Krokan HE: Low copy number DNA template can render polymerase chain reaction error prone in a sequence-dependent manner. J Mol Diagn 2005, 7(1):36-39.
57. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB: Characterizing and measuring bias in sequence data. Genome biology 2013, 14(5):R51.
58. McCarthy A: Third generation DNA sequencing: pacific biosciences' single molecule real time technology. Chemistry & biology 2010, 17(7):675-676.
59. Kircher M, Stenzel U, Kelso J: Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome biology 2009, 10(8).
60. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H et al: Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 2011, 39(13):e90.
61. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW: Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS computational biology 2013, 9(4):e1003031.
62. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R: Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences of the United States of America 2011, 108(50):20166-20171.
63. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL: High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proceedings of the National Academy of Sciences of the United States of America 2013, 110(49):19872-19877.
64. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA et al: Detecting ultralow-frequency mutations by Duplex Sequencing. Nature protocols 2014, 9(11):2586-2606.
65. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011, 43(5):491-+.
66. Milbury CA, Li J, Liu P, Makrigiorgos GM: COLD-PCR: improving the sensitivity of molecular diagnostics assays. Expert review of molecular diagnostics 2011, 11(2):159-169.
67. Li J, Wang L, Mamon H, Kulke MH, Berbeco R, Makrigiorgos GM: Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. Nature medicine 2008, 14(5):579-584.

68. Li J, Makrigiorgos GM: COLD-PCR: a new platform for highly improved mutation detection in cancer and genetic testing. Biochemical Society transactions 2009, 37(Pt 2):427-432.
69. Milbury CA, Correll M, Quackenbush J, Rubio R, Makrigiorgos GM: COLD-PCR enrichment of rare cancer mutations prior to targeted amplicon resequencing. Clinical chemistry 2012, 58(3):580-589.
70. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA: Sequencing small genomic targets with high efficiency and extreme accuracy. Nature methods 2015.
71. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012, 22(3):568-576.
72. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK: Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012, 28(14):1811-1817.
73. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology 2013, 31(3):213-219.
74. Harismendy O, Schwab RB, Bao L, Olson J, Rozenzhak S, Kotsopoulos SK, Pond S, Crain B, Chee MS, Messer K et al: Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol 2011, 12(12):R124.
75. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N: LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 2012, 40(22):11189-11201.
76. Holtgrewe M, Emde AK, Weese D, Reinert K: A novel and well-defined benchmarking method for second generation read mapping. BMC bioinformatics 2011, 12:210.
77. Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N et al: pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics 2012, 28(11):1533-1535.
78. Huang W, Li L, Myers JR, Marth GT: ART: a next-generation sequencing read simulator. Bioinformatics 2012, 28(4):593-594.
79. McElroy KE, Luciani F, Thomas T: GemSIM: general, error-model based simulator of next-generation sequencing data. BMC genomics 2012, 13:74.
80. Pattnaik S, Gupta S, Rao AA, Panda B: SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. BMC bioinformatics 2014, 15:40.
81. Talwalkar A, Liptrap J, Newcomb J, Hartl C, Terhorst J, Curtis K, Bresler M, Song YS, Jordan MI, Patterson D: SMaSH: a benchmarking toolkit for human genome variant calling. Bioinformatics 2014, 30(19):2787-2795.
82. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY et al: Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature methods 2015, 12(7):623-630.
83. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y: Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC genomics 2014, 15:244.

84. Allhoff M, Schonhuth A, Martin M, Costa IG, Rahmann S, Marschall T: Discovering motifs that induce sequencing errors. BMC bioinformatics 2013, 14.
85. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: An integrated map of genetic variation from 1,092 human genomes. Nature 2012, 491(7422):56-65.
86. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA et al: Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012, 30(5):413-421.
87. Mitchell TM: Machine Learning: The Mc-Graw-Hill Companies, Inc.; 1997.
88. Fayyad UMI, Keki B.: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In: Proceedings of the International Joint Conference on Uncertainty in AI 1993; 1993: 1022-1027.
89. Breiman L: Random Forests. Machine Learning 2001, 45(1):5-32.
90. Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian HW: The WEKA data mining software: an update. SIGKDD Explor Newsl 2009, 11(1):10-18.
91. Torrent Variant Caller Parameters [http://mendel.iontorrent.com/ion-docs/Torrent-Variant-Caller-Parameters.html]
92. Somatic variant calling workflow for matched tumor-normal samples [https://sites.google.com/site/strelkasomaticvariantcaller/home]
93. Variant detection in massively parallel sequencing data [http://varscan.sourceforge.net/using-varscan.html]
94. Long JS: Regression Models for Categorical and Limited Dependent Variables: Sage; 1997.
95. McFadden D: Quantitative Methods for Analyzing Travel Behaviour on Individuals: Some Recent Developments. In: Bahvioural Travel Modelling. Edited by Stopher DHaP; 1979: 305.
96. Zimmermann MRVKF: Evaluating Pseudo-R2's for binary probit models. Quality & Quantity 1994(28):151-164.
97. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA: Sequencing small genomic targets with high efficiency and extreme accuracy. Nature methods 2015, 12(5):423-425.
98. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, Dawson SJ, Piskorz AM, Jimenez-Linan M, Bentley D et al: Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Science translational medicine 2012, 4(136):136ra168.
99. McGranahan N, Swanton C: Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell 2015, 27(1):15-26.
100. Michor F, Polyak K: The origins and implications of intratumor heterogeneity. Cancer prevention research 2010, 3(11):1361-1364.
101. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P et al: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine 2012, 366(10):883-892.
102. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H et al: Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nature medicine 2015, 21(7):751-759.
103. Suprynowicz FA, Upadhyay G, Krawczyk E, Kramer SC, Hebert JD, Liu X, Yuan H, Cheluvaraju C, Clapp PW, Boucher RC, Jr. et al: Conditionally reprogrammed cells represent a stem-like state of adult epithelial cells. Proceedings of the

National Academy of Sciences of the United States of America 2012, 109(49):20035-20040.
104. Liu X, Ory V, Chapman S, Yuan H, Albanese C, Kallakury B, Timofeeva OA, Nealon C, Dakic A, Simic V et al: ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. The American journal of pathology 2012, 180(2):599-607.
105. Nakshatri H, Anjanappa M, Bhat-Nakshatri P: Ethnicity-Dependent and -Independent Heterogeneity in Healthy Normal Breast Hierarchy Impacts Tumor Characterization. Scientific reports 2015, 5:13526.
106. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. Genome Res 2002, 12(6):996-1006.
107. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al: The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 2015, 43(D1):D670-D681.
108. Emes RD, Goodstadt L, Winter EE, Ponting CP: Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet 2003, 12(7):701-709.
109. Why Mouse Matters [https://http://www.genome.gov/10001345]
110. Jane-Valbuena J, Widlund HR, Perner S, Johnson LA, Dibner AC, Lin WM, Baker AC, Nazarian RM, Vijayendran KG, Sellers WR et al: An Oncogenic Role for ETV1 in Melanoma. Cancer Res 2010, 70(5):2075-2084.
111. Oh S, Shin S, Janknecht R: ETV1, 4 and 5: An oncogenic subfamily of ETS transcription factors. Bba-Rev Cancer 2012, 1826(1):1-12.
112. Coutte L, Monte D, Imai K, Pouilly L, Dewitte F, Vidaud M, Adamski J, Baert JL, de Launoit Y: Characterization of the human and mouse ETV1/ER81 transcription factor genes: role of the two alternatively spliced isoforms in the human. Oncogene 1999, 18(46):6278-6286.
113. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao XH, Morris DS, Menon A, Jing XJ, Cao Q, Han B et al: Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007, 448(7153):595-U599.
114. Guillon N, Tirode F, Boeva V, Zynovyev A, Barillot E, Delattre O: The Oncogenic EWS-FLI1 Protein Binds In Vivo GGAA Microsatellite Sequences with Potential Transcriptional Activation Function. PloS one 2009, 4(3).
115. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC et al: Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature 2012, 486(7403):353-360.
116. Liang S, He L, Zhao X, Miao Y, Gu Y, Guo C, Xue Z, Dou W, Hu F, Wu K et al: MicroRNA let-7f inhibits tumor invasion and metastasis by targeting MYH9 in human gastric cancer. PloS one 2011, 6(4):e18409.
117. Bader AG, Kang SY, Vogt PK: Cancer-specific mutations in PIK3CA are oncogenic in vivo. Proceedings of the National Academy of Sciences of the United States of America 2006, 103(5):1475-1479.
118. Karakas B, Bachman KE, Park BH: Mutation of the PIK3CA oncogene in human cancers. Brit J Cancer 2006, 94(4):455-459.
119. Miller TW: Initiating breast cancer by PIK3CA mutation. Breast Cancer Res 2012, 14(1).
120. Janku F, Wheler JJ, Naing A, Falchook GS, Hong DS, Stepanek VM, Fu SQ, Piha-Paul SA, Lee JJ, Luthra R et al: PIK3CA Mutation H1047R Is Associated with Response to PI3K/AKT/mTOR Signaling Pathway Inhibitors in Early-Phase Clinical Trials. Cancer Res 2013, 73(1):276-284.

121. Meyer DS, Koren S, Leroy C, Brinkhaus H, Muller U, Klebba I, Muller M, Cardiff RD, Bentires-Alj M: Expression of PIK3CA mutant E545K in the mammary gland induces heterogeneous tumors but is less potent than mutant H1047R. Oncogenesis 2013, 2.
122. Gkeka P, Evangelidis T, Pavlaki M, Lazani V, Christoforidis S, Agianian B, Cournia Z: Investigating the Structure and Dynamics of the PIK3CA Wild-Type and H1047R Oncogenic Mutant. PLoS computational biology 2014, 10(10).
123. Bouchard C, Lee S, Paulus-Hock V, Loddenkemper C, Eilers M, Schmitt CA: FoxO transcription factors suppress Myc-driven lymphomagenesis via direct activation of Arf. Gene Dev 2007, 21(21):2775-2787.
124. Fu Z, Tindall DJ: FOXOs, cancer and regulation of apoptosis. Oncogene 2008, 27(16):2312-2319.
125. Kloet DEA, Burgering BMT: The PKB/FOXO switch in aging and cancer. Bba-Mol Cell Res 2011, 1813(11):1926-1937.
126. Zhang XB, Tang NM, Hadden TJ, Rishi AK: Akt, FoxO and regulation of apoptosis. Bba-Mol Cell Res 2011, 1813(11):1978-1986.
127. Diep CH, Charles NJ, Gilks CB, Kalloger SE, Argenta PA, Lange CA: Progesterone receptors induce FOXO1-dependent senescence in ovarian cancer cells. Cell Cycle 2013, 12(9):1433-1449.
128. Zhang P, Tu B, Wang H, Cao ZY, Tang M, Zhang CH, Gu B, Li ZM, Wang LN, Yang Y et al: Tumor suppressor p53 cooperates with SIRT6 to regulate gluconeogenesis by promoting FoxO1 nuclear exclusion. Proceedings of the National Academy of Sciences of the United States of America 2014, 111(29):10684-10689.
129. Cheung LWT, Hennessy BT, Li J, Yu SX, Myers AP, Djordjevic B, Lu YL, Stemke-Hale K, Dyer MD, Zhang F et al: High Frequency of PIK3R1 and PIK3R2 Mutations in Endometrial Cancer Elucidates a Novel Mechanism for Regulation of PTEN Protein Stability. Cancer Discov 2011, 1(2):170-185.
130. Urick ME, Rudd ML, Godwin AK, Sgroi D, Merino M, Bell DW: PIK3R1 (p85 alpha) Is Somatically Mutated at High Frequency in Primary Endometrial Cancer. Cancer Res 2011, 71(12):4061-4067.
131. Cheung LWT: High Frequency of PIK3R1 and PIK3R2 Mutations in Endometrial Cancer Elucidates a Novel Mechanism for Regulation of PTEN Protein Stability. Cancer Discov 2012, 2(8):750-751.
132. Cheung LWT, Yu SX, Zhang D, Li J, Ng PKS, Panupinthu N, Mitra S, Ju ZL, Yu QH, Liang H et al: Naturally Occurring Neomorphic PIK3R1 Mutations Activate the MAPK Pathway, Dictating Therapeutic Response to MAPK Pathway Inhibitors. Cancer Cell 2014, 26(4):479-494.
133. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA: The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data (vol 2, pg 401, 2012). Cancer Discov 2012, 2(10):960-960.
134. Gao JJ, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun YC, Jacobsen A, Sinha R, Larsson E et al: Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci Signal 2013, 6(269).
135. The I, Ruijtenberg S, Bouchet BP, Cristobal A, Prinsen MB, van Mourik T, Koreth J, Xu H, Heck AJ, Akhmanova A et al: Rb and FZR1/Cdh1 determine CDK4/6-cyclin D requirement in C. elegans and human cancer cells. Nature communications 2015, 6:5906.

136. Puto LA, Reed JC: Daxx represses RelB target promoters via DNA methyltransferase recruitment and DNA hypermethylation. Gene Dev 2008, 22(8):998-1010.
137. Lewis PW, Elsaesser SJ, Noh KM, Stadler SC, Allis CD: Daxx is an H3.3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres. Proceedings of the National Academy of Sciences of the United States of America 2010, 107(32):14075-14080.
138. Rapkin LM, Ahmed K, Dulev S, Li R, Kimura H, Ishov AM, Bazett-Jones DP: The histone chaperone DAXX maintains the structural organization of heterochromatin domains. Epigenet Chromatin 2015, 8.
139. Shalginskikh N, Poleshko A, Skalka AM, Katz RA: Retroviral DNA Methylation and Epigenetic Repression Are Mediated by the Antiviral Host Protein Daxx. J Virol 2013, 87(4):2137-2150.
140. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD et al: Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012, 481(7382):506-510.
141. Hoaglin DC: A poissonness plot. The American Statistician 1980, 34, No.3:146-149.
142. David C. Hoaglin FMaJWT: Checking the Shape of Discrete Distributions. In: Checking the Shape of Discrete Distributions, in Exploring Data Tables, Trends, and Shapes. Edited by David C. Hoaglin FMaJWT. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2011.
143. Friendly M: Visualizing Categorical Data. Cary, NC: SAS Institute; 2000.
144. David Meyer AZ, and Kurt Hornik: vcd: Visualizing Categorical Data. In., 1.4-1 edn; 2015.
145. Hoaglin DCFMJWT: Exploring Data Tables, Trends and Shapes.: Wiley; 1985.
146. Benjamini Y, Hochberg Y: Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 1995, 57(1):289-300.
147. Lambert D: Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics 1992, 34(1):1-14.
148. Greene WH: Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. In.: NYU Working Paper No. EC-94-10; 1994.
149. Vuong QH: Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 1989, 57(2):307-333.
150. Desmarais BA, Harden JJ: Testing for zero inflation in count models: Bias correction for the Vuong test. Stata J 2013, 13(4):810-835.
151. Cameron AC, Trivedi PK: Regression-Based Tests for Overdispersion in the Poisson Model. J Econometrics 1990, 46(3):347-364.
152. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26(1):139-140.
153. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, Hutchison JB, Larson JW et al: Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 2009, 27(11):1025-1031.
154. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C: The sequence read archive. Nucleic Acids Res 2011, 39(Database issue):D19-21.

155. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C: The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012, 40(Database issue):D54-56.
156. Minoche AE, Dohm JC, Himmelbauer H: Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome biology 2011, 12(11):R112.
157. Anders S, Huber W: Differential expression analysis for sequence count data. Genome biology 2010, 11(10):R106.

CURRICULUM VITAE

Yangyang Hao

Education

08/2010 – 02/2016	Indiana University, IUPUI, Indianapolis, IN Major: Medical & Molecular Genetics, Doctor of Philosophy Minor: Life Science
08/2007 – 07/2009	Harbin Institute of Technology, Harbin, Heilongjiang, China Major: Computer Science, Master of Engineering Minor: Bioinformatics
08/2003 – 07/2007	Harbin Institute of Technology, Harbin, Heilongjiang, China Major: Computer Science, Bachelor of Engineering Minor: Bioinformatics

Honors, Awards and Fellowships

08/2010 – 09/2015	Indiana University, IUPUI, Indianapolis, IN School of Medicine Fellowship toward a Ph.D. Degree in Medical & Molecular Genetics
07/2009	Harbin Institute of Technology, Harbin, Heilongjiang, China Outstanding Master Graduate in Department of Computer Science and Engineering
2007, 2008	Harbin Institute of Technology, Harbin, Heilongjiang, China Top Class Scholarship for Postgraduates
2006, 2007, 2008	Harbin Institute of Technology, Harbin, Heilongjiang, China Excellent Student Leader
2006	Harbin Institute of Technology, Harbin, Heilongjiang, China National Endeavor Scholarship
2005	Harbin Institute of Technology, Harbin, Heilongjiang, China Fuji Xerox Scholarship

Research, Teaching and Work Experience

12/2015 – now	Veracyte, South San Francisco, CA Bioinformatics Scientist
09/2015 – 12/2015	Veracyte, South San Francisco, CA Bioinformatics Scientist Intern
08/2010 – 11/2015	Indiana University, IUPUI, Indianapolis, IN

	Department of Medical & Molecular Genetics Research Assistant in Prof. Yunlong Liu's lab
08/2014 – 12/2014	Indiana University, IUPUI, Indianapolis, IN Department of Medical & Molecular Genetics Teaching Assistant for G788 Introduction to Next Generation Sequencing
08/2007 – 07/2010	Harbin Institute of Technology, Harbin, Heilongjiang, China Department of Computer Science and Engineering Bioinformatics Research Assistant in Prof. Yadong Wang's lab
03/2009 – 06/2009	Harbin Institute of Technology, Harbin, Heilongjiang, China Department of Computer Science and Engineering Teaching Assistant for Genomic Informatics
09/2008 – 12/2008	Harbin Institute of Technology, Harbin, Heilongjiang, China Department of Computer Science and Engineering Teaching Assistant for Computer Organization Principle
04/2008 – 06/2008	Harbin Institute of Technology, Harbin, Heilongjiang, China Department of Computer Science and Engineering Teaching Assistant for Data Structure
09/2007 – 12/2007	Harbin Institute of Technology, Harbin, Heilongjiang, China Department of Computer Science and Engineering Teaching Assistant for Pattern Reorganization

Conferences Attended

01/10/2016	Asia Pacific Bioinformatics Conference (APBC) 2016, South San Francisco, CA Short Courses
11/2015	International Conference on Intelligent Biology and Medicine (ICIBM) 2015 Indiana University, IUPUI, Indianapolis, IN Talk on "Statistical modeling for sensitive detection of low-frequency single nucleotide variants"
09/2012	ICBP/PS-OC Junior Investigator Meeting 2012 Seattle, WA
09/2011	Statistical Analyses for Next Generation Sequencing 2011 The University of Alabama at Birmingham, Birmingham, AL Poster "Statistical Modeling for Pooled Sequencing Variants Identification"

Journal Publications

Xinjun Zhang, Hai Lin, Huiying Zhao, **Yangyang Hao**, Matthew Mort, David N Cooper, Yaoqi Zhou, Yunlong Liu: Impact of Human Pathogenic Micro-Insertions and Micro-Deletions on Post-Transcriptional Regulation. Human Molecular Genetics 01/2014.
Ao Zhou, Marcus R Breese, **Yangyang Hao**, Howard J Edenberg, Lang Li, Todd C Skaar, Yunlong Liu: Alt Event Finder: a tool for extracting alternative splicing events from RNA-seq data. BMC Genomics 12/2012.
Qinghua Jiang*, **Yangyang Hao***, Guohua Wang, Liran Juan, Tianjiao Zhang, Mingxiang Teng, Yunlong Liu, Yadong Wang: Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Systems Biology 01/2010; 4 Suppl 1:S2.
Qinghua Jiang, Yadong Wang, **Yangyang Hao**, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, Yunlong Liu: miR2Disease: a manually curated database for microRNA deregulation in human disease.. Nucleic Acids Research 11/2008; 37(Database issue): D98-104.

Conference Proceedings

Qinghua Jiang, **Yangyang Hao**, Guohua Wang, Tianjiao Zhang, Yadong Wang: Weighted Network-Based Inference of Human MicroRNA-Disease Associations. Fifth International Conference on Frontier of Computer Science and Technology, FCST 2010, Changchun, Jilin Province, China, August 18-22, 2010; 01/2010.

Manuscripts Accepted

Yangyang Hao, Pengyue Zhang, Xiaoling Xuei, Harikrishna Nakshatri, Howard Edenberg, Lang Li, Yunlong Liu: Statistical modeling for sensitive detection of low-frequency single nucleotide variants. BMC Genomics Supplement Issue for ICIBM 2015, 1471-2164-17-S1-13.
Weixing Feng, Dingkai Xue, Fengfei Song, Sen Zhao, Ziwei Li, Duojiao Chen, **Yangyang Hao** and Yunlong Liu. Improving alignment accuracy on homopolymer regions for semiconductor based sequencing technologies. BMC Genomics Supplement Issue for ICIBM 2015, 1471-2164-17-S1-43.

Conference Abstracts

Manjushree Anjanappa, **Yangyang Hao**, Howard J Edenberg, Yunlong Liu, and Harikrishna Nakshatri. Identification of cancer-specific signaling networks: what is “normal” control? (AACR 2016 Meeting Abstract.)
Poornima Bhat-Nakshatri, Manjushree Anjanappa, **Yangyang Hao**, Hitesh N.Appaiah1, Howard Edenberg, Yunlong Liu, Harikrishna Nakshatri: Estradiol-inducible Dependence Receptor UNC5a restricts Estrogen Receptor Activity and Imparts Estradiol Dependence to Breast Cancer Cells. (AACR 2015 Meeting Abstract.)

Manuscripts in Preparation

Yangyang Hao, Xiaoling Xuei, Howard Edenberg, Lang Li, Harikrishna Nakshatri, Yunlong Liu: RareVar: a Framework for Detecting Low Frequency Single Nucleotide Variants.

Yangyang Hao, Manjushree Anjanappa, Xiaoling Xuei, Howard Edenberg, Yunlong Liu, Harikrishna Nakshatri: Breast Cancer Cell Line Enabled Rare SNPs Targeting Drug Selection for DCIS and LCIS.

Li Qin, **Yangyang Hao**, Yunlong Liu, Jianting Zhang: Reversible methylation of PDGFD by next-generation gene sequencing and its essential role in acquired gemcitabine resistance.

Derek P. Logsdon, Huiwen Cheng, **Yangyang Hao**, Safi Shahda, Yanlin Jiang, Mircea Ivan, Cecilia Devlin, Yunlong Liu, Mark R. Kelley, Melissa L. Fishel: Targeting APE-1/Ref-1 Results in Inhibition of Hypoxia Signaling Genes.