Briefings in Bioinformatics Advance Access published September 24, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 11 doi:10.1093/bib/bbt066

Three-stage quality control strategies for DNA re-sequencing data

Yan Guo, Fei Ye, Quanghu Sheng, Travis Clark and David C. Samuels

Submitted: 13th June 2013; Received (in revised form): 21st August 2013

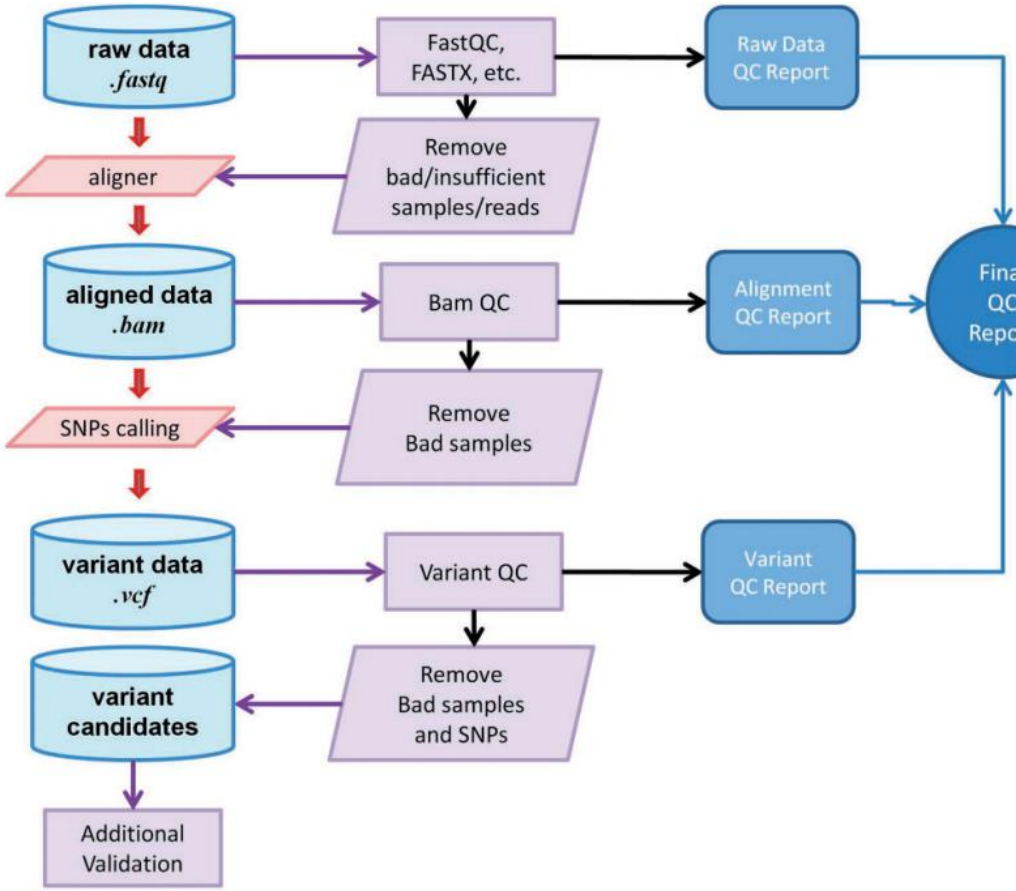


Figure 4: Overall workflow of quality control in DNA sequencing data.

Quality control of the raw data

① base quality: 文中给出Illumina的base quality数个特点；
② the nucleotide distribution:
③ GC content distribution: genome(49%-51%) while exome(38%-39%);
④ the duplication rate: 不同lane的同一个sample被测reads相差多大?

Alignment quality control

① Different alignment quality control parameters should be collected for exome and whole-genome sequencing.
② The most important quality control parameter for whole-genome sequencing is the average or median depth and the percentage of the genome covered by the sequencing at that depth.
③ Capture efficiency is the most important quality control for exome sequencing or other targeted sequencing.

Quality control on variant calling

① For human genome, the Ti/Tv ratio is around 3.0 for SNPs inside exons and aabout 2.0 elsewhere, and the ratio also differs between synonymous and non-synonymous SNPs.
② A higher number (200) of novel non-synonymous SNPs would likely indicate a high false-positive.
③ GATk variant quality score recalibration.
④ The heterozygosity to non-reference homozygosity ratio is 2 for whole-genome data.
⑤ Commonly used base quality score threshold for BWA is 20. Draw the distribution of mapping quality scores and examine this distribution for outliers.
⑥ Base alignment quality (BAQ) reduces the FP calls by decreasing the base quality scores for base around insertion and deletion events in the sequence.(不能和GATK的local realignment同时使用，不然会引入FP calling)
⑦ A set of filters recommended by GATK
⑧ Strand bias
⑨ Reference allele preferential
⑩ SNP density (two SNPs within 10 bp)

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2498–2500
doi:10.1093/bioinformatics/btt422

Sequence analysis

Advance Access publication August 13, 2013

DRAW+SneakPeek: Analysis workflow and quality metric management for DNA-seq experiments

Chiao-Feng Lin[1,2], Otto Valladares[1,2], D. Micah Childress[1,2], Egor Klevak[3], Evan T. Geller[1], Yih-Chii Hwang[2,4], Ellen A. Tsai[4,5], Gerard D. Schellenberg[1,*] and Li-San Wang[1,2,*]

[1]Department of Pathology and Laboratory Medicine and [2]Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, [3]Department of Physics, University of Washington, Seattle, WA 98105, USA, [4]Genomics and Computational Biology Graduate Group, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA and [5]Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Associate Editor: Martin Bishop

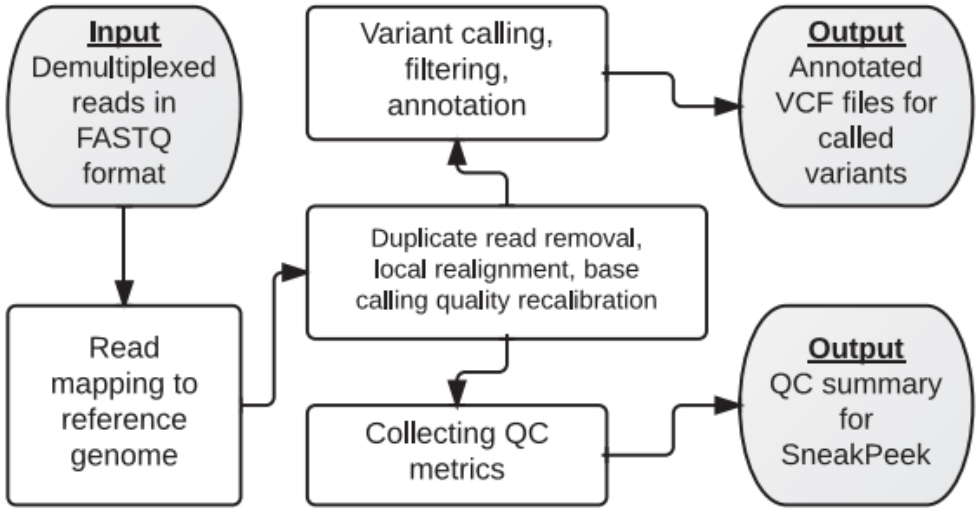
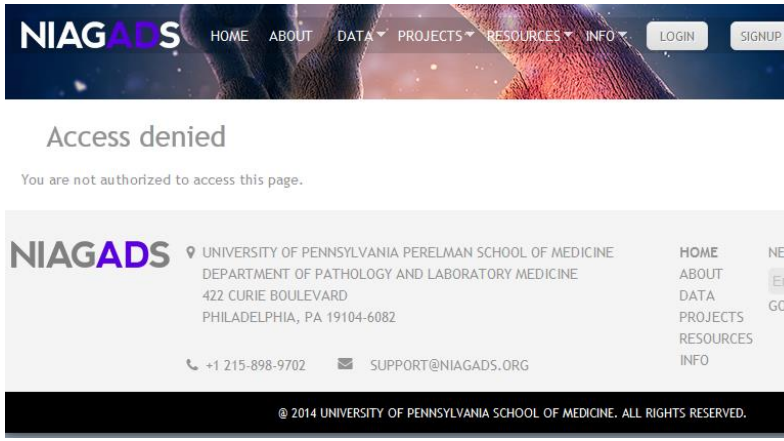


Fig. 1. DRAW+SneakPeek overview

① 分两部分，第一部分DRAW用的是GATK的流程，第二部分通过把variant的结果放入数据库(mysql)， SneakPeek形成可视化界面，便于比较。**SneakPeek is a quality metrics management system for reviewing the sequencing qualtiy of multiple samples across different flow cells.**

② 在亚马孙云上有image，我现在没有账号，没试过。根据给出的网站去看，说没有浏览权限。



BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1767–1768
doi:10.1093/bioinformatics/btu108

Sequence analysis

Advance Access publication February 26, 2014

ngsCAT: a tool to assess the efficiency of targeted enrichment sequencing

Francisco J. López-Domingo[1], Javier P. Florido[1], Antonio Rueda[1], Joaquín Dopazo[1,2,3] and Javier Santoyo-Lopez[1,*]

[1]Bioinformatics Department, Genomics and Bioinformatics Platform of Andalusia (GBPA), 41092 Seville, [2]Computational Genomics Department, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain and [3]Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain

Associate Editor: Inanc Birol

Requirements

① Operating system: Linux-based OS
② Computer processor: multi-core processor
③ Computer memory: 4GB or more

Dependencies

① samtools
② numpy: a python package for scientific computing with Python.
③ scipy: a python library which provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization.
④ xlwt: a python library for generating spreadsheet files that are compatible with Excel 97/2000/XP/2003, OpenOffice.org Calc, and Gnumeric.
⑤ matplotlib: a python library which produces high-quality 2D graphics for interactive graphing, scientific publishing, user interface development and web application servers targeting multiple user interfaces and hardcopy output formats.
⑥ pysam: a python module for reading and manipulating Sam/Bam files.

BAM file + BED file ← 输入

↓

ngsCAT

↓

输出

① **the number and percentage of reads on/off target**
② **the percentage of target bases covered at different coverage thresholds**
③ **the number of duplicated reads on/off target**
④ **bedgraph tracks of off-target regions with high coverage**
⑤ **the distribution of the coverage in the ROIs**
⑥ **the variability of the coverage within the ROIs**
⑦ **the distribution of the coverage as a function of GC content**
⑧ **a satur-ation curve of the coverage as a function of the number of reads that can serve to estimate whether sequencing a higher number of reads will produce a significant increase of the coverage in the ROIs**
⑨ **process two samples at a time, which allows a simple comparison of two samples**
⑩ **analysis of small target regions as well as for larger regions like whole exomes**

http://www.bioinfomgp.org/ngscat

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 3 2014, pages 442–444
doi:10.1093/bioinformatics/btt682

Genetics and population analysis

Advance Access publication December 12, 2013

PhenoMan: phenotypic data exploration, selection, management and quality control for association studies of rare and common variants

Biao Li, Gao Wang and Suzanne M. Leal*

Center for Statistical Genetics, Department of Molecular and Human Genetics, One Baylor Plaza 700D, Baylor College of Medicine, Houston, TX 77030, USA

Associate Editor: Jeffrey Barrett

1. 这个工具主要是用于表型的quality control。
2. PhenoMan, an interactive command-line driven program, is written in Python and R.
3. 工具可以实现的内容：

① 通过各种性状或数量性状选择样本；
② 过滤不符合的样本；(例如不属于同一人种，或性别不对等)
③ 重新编码各种变量；例如缺失的变量用平均值补上；
④ 对数量性状的值做各种变换，例如log变换；
⑤ 选择covariates(?)；
⑥ 对数据做总结；

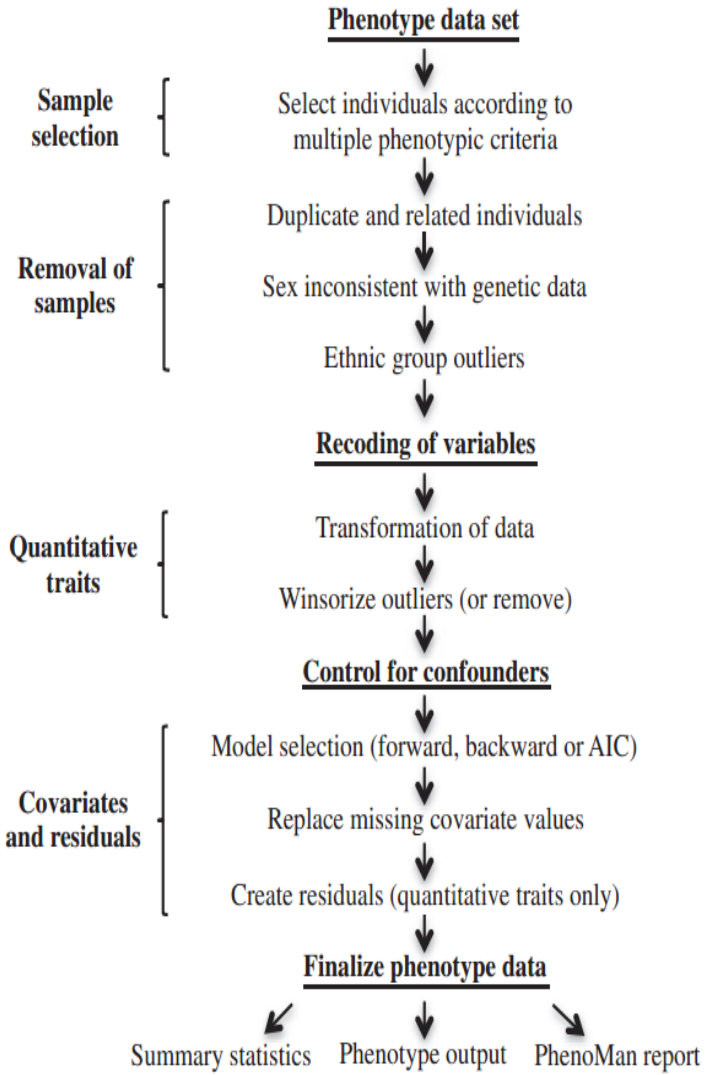


Fig. 1. Schematic workflow for PhenoMan

https://code.google.com/p/phenoman/

Emma_JiangChongyi

SeqControl: process control for DNA sequencing

Lauren C Chong[1], Marco A Albuquerque[1], Nicholas J Harding[1], Cristian Caloian[1], Michelle Chan-Seng-Yue[1], Richard de Borja[1], Michael Fraser[2], Robert E Denroche[1], Timothy A Beck[1], Theodorus van der Kwast[2], Robert G Bristow[3,4], John D McPherson[4,5] & Paul C Boutros[1,4,6]

As high-throughput sequencing continues to increase in speed and throughput, routine clinical and industrial application draws closer. These 'production' settings will require enhanced quality monitoring and quality control to optimize output and reduce costs. We developed SeqControl, a framework for predicting sequencing quality and coverage using a set of 15 metrics describing overall coverage, coverage distribution, basewise coverage and basewise quality. Using whole-genome sequences of 27 prostate cancers and 26 normal references, we derived multivariate models that predict sequencing quality and depth. SeqControl robustly predicted how much sequencing was required to reach a given coverage depth (area under the curve (AUC) = 0.993), accurately classified clinically relevant formalin-fixed, paraffin-embedded samples, and made predictions from as little as one-eighth of a sequencing lane (AUC = 0.967). These techniques can be immediately incorporated into existing sequencing pipelines to monitor data quality in real time. SeqControl is available at http://labs.oicr.on.ca/Boutros-lab/software/SeqControl/.

or delays for additional data collection. Groups developing and evaluating new protocols must rapidly assess whether these methods improve or degrade data quality.

To date, such techniques have been elusive, with only a handful of heuristic studies performed[19–23]. Several factors confound prediction of sequencing quality, including variability among machines and reagent batches as well as the complexity of sequencing libraries[23] (**Supplementary Note**). Similarly, the integrity and quality of DNA vary widely: formalin-fixed, paraffin-embedded (FFPE) samples typically yield degraded DNA that is challenging to sequence[24,25]. With large-scale exome and whole-genome sequencing studies[26] increasing in prevalence, the need for robust quality control is growing. To tackle these challenges we introduce SeqControl: a technique for predicting overall experimental qualities using a small amount of sequencing, enabling production-scale use of sequencing quality control and process control. Using 15 quality metrics evaluated on whole-genome sequencing data from prostate tumors and matched blood samples, we predicted whether experiments would achieve their target sequencing depth using as little as one-eighth

这篇文献只明白80%左右

Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook[1], Brad Chapman[2], Jason Wang[3], David Mittelman[3,4], Oliver Hofmann[2], Winston Hide[2] & Marc Salit[1]

Clinical adoption of human genome sequencing requires methods that output genotypes with known accuracy at millions or billions of positions across a genome. Because of substantial discordance among calls made by existing sequencing methods and algorithms, there is a need for a highly accurate set of genotypes across a genome that can be used as a benchmark. Here we present methods to make high-confidence, single-nucleotide polymorphism (SNP), indel and homozygous reference genotype calls for NA12878, the pilot genome for the Genome in a Bottle Consortium. We minimize bias toward any method by integrating and arbitrating between 14 data sets from five sequencing technologies, seven read mappers and three variant callers. We identify regions for which no confident genotype call could be made, and classify them into different categories based on reasons for uncertainty. Our genotype calls are publicly available on the Genome Comparison and Analytic Testing website to enable real-time benchmarking of

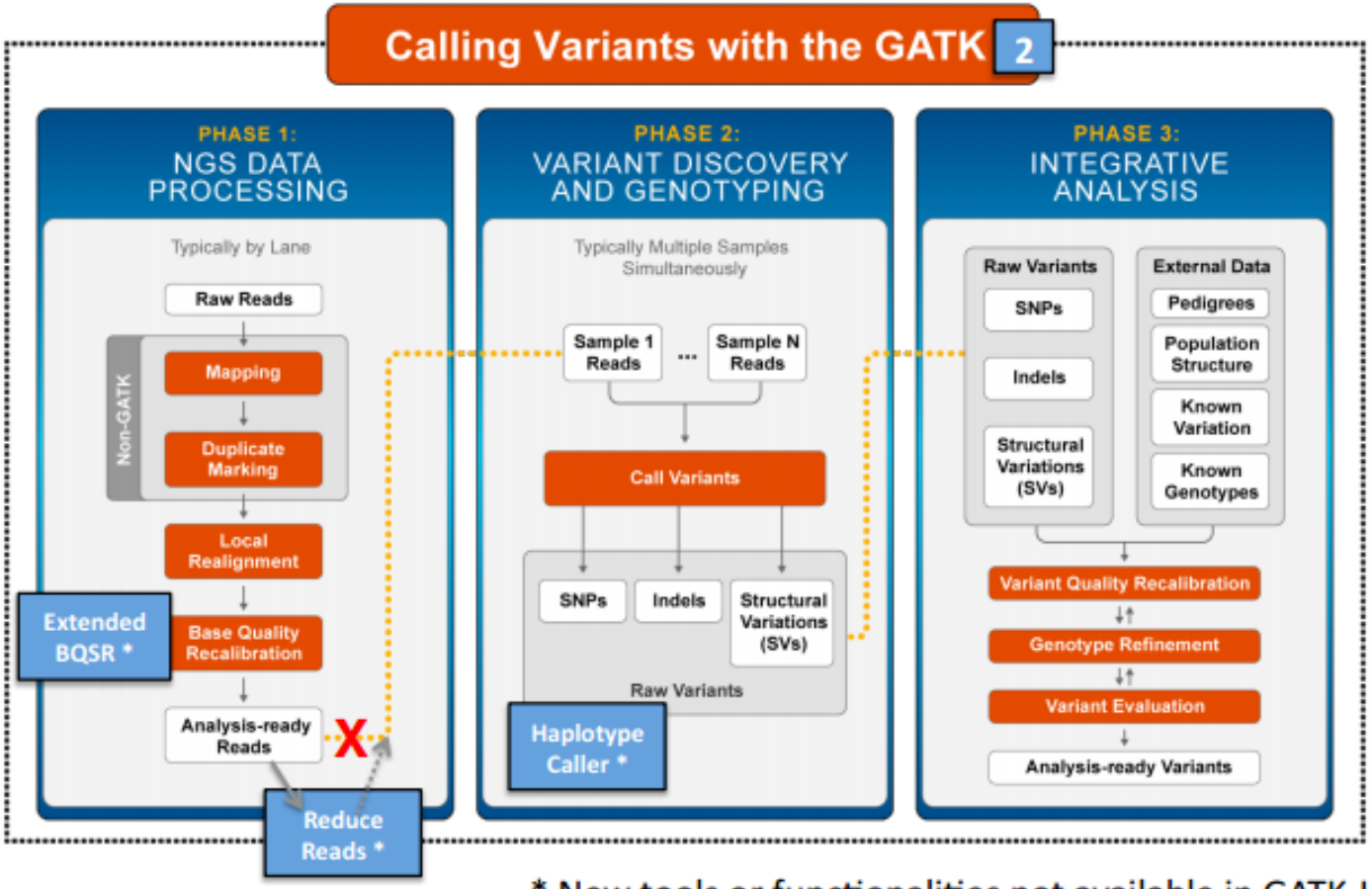
Administration highlighted the utility of this candidate NIST reference material in approving the assay for clinical use[12].

NIST, with the Genome in a Bottle Consortium, is developing well-characterized whole-genome reference materials, which will be available to research, commercial and clinical laboratories for sequencing and assessing variant-call accuracy and understanding biases. The creation of whole-genome reference materials requires a best estimate of what is in each tube of DNA reference material, describing potential biases and estimating the confidence of the reported characteristics. To develop these data, we are developing methods to arbitrate between results from multiple sequencing and bioinformatics methods. The resulting arbitrated integrated genotypes can then be used as a benchmark to assess rates of false positives (o r calling a variant at a homozygous reference site), false negatives (or calling homozygous reference at a variant site) and other genotype calling errors (e.g., calling homozygous variant at a heterozygous site).

Current methods for assessing sequencing performance are

这篇文献还没有看完，尚未深入解读。

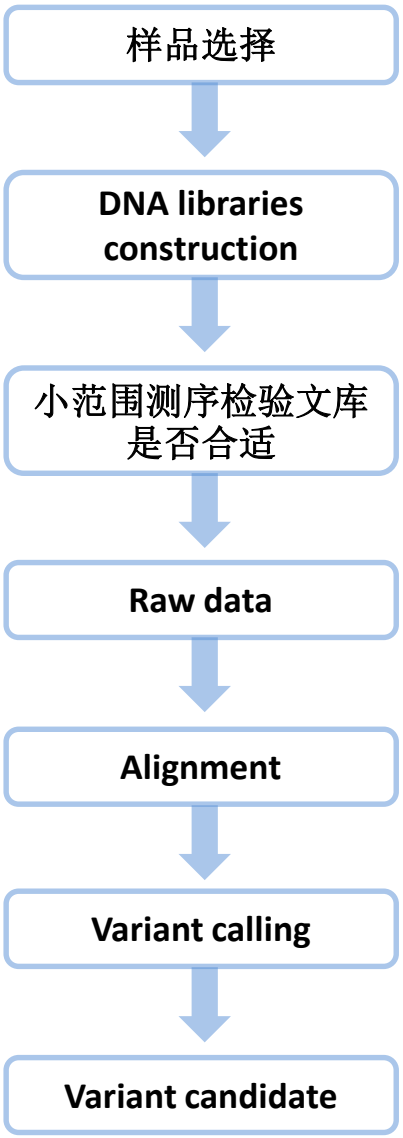
Scope and schema of the Best Practices



* New tools or functionalities not available in GATK-Lite

根据GATK官网的教程对其3个Phase的每一步进行学习

根据这段时间的文献调研(看懂的，没看懂的，还没看完的)，我NGS QC流程了解如下：



PhenoMan (这个工具是不是可以在建样品数据库的时候用到？)

Exome的library在捕获前需要confirmation of size by running on Agilent Bioanalyzer 2100 High Sensitibity Chip.

SeqControl
Measure important features of the library

FastQC或者FastX (4个方面)

选用合适的reference，见文献[1]; BWA 或者自带软件
Average or median depth and the percentage of the genome covered by the sequencing at that depth

GATK

Ti/Tv, number of novel non-synonymous SNPs, heterozygosity to non-reference homozygosity ratio …

文献[1] …rating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

OPEN ACCESS Freely available online

PLOS ONE

Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer

Zhen Xuan Yeo[1], Maurice Chan[1], Yoon Sim Yap[2], Peter Ang[2,3], Steve Rozen[4], Ann Siew Gek Lee[1]*

1 Division of Medical Sciences, National Cancer Centre Singapore, Singapore, Singapore, **2** Department of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore, **3** Oncocare Cancer Centre, Gleneagles Medical Centre, Singapore, Singapore, **4** Neuroscience and Behavioral Disorders, Duke-NUS Graduate Medical School, Singapore, Singapore

① Homopolymer sequencing errors are those associated with runs of consecutive identical nucleotides. These errors tend to increase in genomics regions where the occurrence of true polymorphisms is also high.

② 前面的文献调研发现Ion Torrent在homopolymer的地方更容易出现single-base deletion。

Table 1. Variants in DH10B detected by different platforms and filter settings. 这个表展示DH10B的genome数据

Sequencer (Variant caller)	Aligner/Workflow	SNVs	Indels	Aligner/Workflow	SNVs	Indels
MiSeq (UnifiedGenotyper)	CASAVA	0**	1**	BWA	0**	1**
SOLiD4 (UnifiedGenotyper)	Bioscope	0**	0	BWA	9	2
PGM (SAMtools)	Torrent Suite	0	42	BWA	0	24
PGM (Dindel)	Torrent Suite	0	478	BWA	0	314
PGM (UnifiedGenotyper)	Torrent Suite	2	204	BWA	0	144
PGM (UnifiedGenotyper + Filtered*)	Torrent Suite	0	1	BWA	0	1

*Filtered using BAF and VARW thresholds.
**Expected number of SNVs and indels.
doi:10.1371/journal.pone.0045798.t001

Overall, when the same data was aligned using BWA, fewer indels were called.

Table 2. Variants in *BRCA1* and *BRCA2* detected by different platforms and filter settings.

The BRCA genes have multiple regions with homopolymer runs, making them ideal candidate to study homopolymers associated errors.

		SNVs				Indels			
Sequencer (Variant caller)	Aligner/Workflow	FP	FN	TP	TN	FP	FN	TP	TN
SOLiD4 (UnifiedGenotyper)	CLC	16	0	33	148744	2	0	3	148788
SOLiD4 (UnifiedGenotyper)	BWA	16	0	33	148744	4	0	3	148786
PGM (SAMtools)	BWA	24	17	16	148736	64	1	2	148726
PGM (UnifiedGenotyper)	BWA	20	0	33	148740	2000	0	3	146790
PGM (UnifiedGenotyper + Filtered*)	BWA	17	0	33	148743	8	0	3	148782

Only 'callable' base were considered which were the sum of all bases in coding exons with ≥4X read coverage from the six samples (n = 148793).
FP = False positives; FN = False negatives; TP = True positives; TN = True negatives, as determined by Sanger sequencing.
*SNVs were filtered using $BAF_{th} = 0.2$. Indels were filtered using $BAF_{th} = 0.28$ and $VARW_{th} = 0$.
doi:10.1371/journal.pone.0045798.t002

综合考虑后还是用UnifiedGenotyper

Yeo et al. BMC Genomics 2014, 15:516
http://www.biomedcentral.com/1471-2164/15/516

BMC Genomics

METHODOLOGY ARTICLE Open Access

Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the *BRCA1* and *BRCA2* genes

Zhen Xuan Yeo[1], Joshua Chee Leong Wong[1], Steven G Rozen[2*] and Ann Siew Gek Lee[1,3,4*]

这篇文献 (2014) 和前一篇文献 (2012) 的作者是同一批人

注意这篇文献只是针对两个基因做的研究。

Table 1 Comparison of indel calling in the 6 training samples using different variant calling workflows, without subsequent filtering

Read mapper	Variant caller	FP[a]	FN[a]	TP[a]	TN[a]	Sensitivity [95% CI]	Specificity [95% CI]	FDR [95% CI]
TMAP-TS2.0	TSVC2.0	0	2	1	96135	33.33% [3.87, 82.33]	100% [100, 100]	0% [0, 77.15]
TMAP-TS2.2	TSVC2.2	0	2	1	96135	33.33% [3.87, 82.33]	100% [100, 100]	0% [0, 77.15]
TMAP-TS3.4	TSVC3.4	8	1	2	96127	66.67% [17.67, 96.13]	99.99% [99.98, 100]	80% [49.72, 95.59]
TMAP-TS2.0	GATK	4	1	2	96131	66.67% [17.67, 96.13]	99.99% [99.99, 100]	66.67% [28.64, 92.32]
TMAP-TS2.2	GATK	9	1	2	96126	66.67% [17.67, 96.13]	99.99% [99.98, 100]	81.82% [53.28, 96.02]
*TMAP-TS3.4	GATK	5	0	3	96130	100% [55.59, 100]	99.99% [99.99, 100]	62.5% [29.48, 88.1]
TMAP-TS2.0	SAMtools	0	3	0	96135	100% [55.59, 100]	99.97% [99.96, 99.98]	90.62% [77.05, 97.29]
TMAP-TS2.2	SAMtools	39	3	0	96096	100% [55.59, 100]	99.99% [99.98, 99.99]	81.25% [57.92, 94.42]
*TMAP-TS3.4	SAMtools	17	0	3	96118	100% [55.59, 100]	99.98% [99.97, 99.99]	85% [65.14, 95.59]
*BWA	GATK	1	0	3	96134	100% [55.59, 100]	99.99% [99.99, 100]	25% [2.85, 71.62]
*BWA	SAMtools	20	0	3	96115	100% [55.59, 100]	99.98% [99.97, 99.99]	86.96% [69.13, 96.19]

We considered all bases in coding exons. Across the 6 samples the total number of bases considered was 96,138.
[a]FP = False Positives; FN = False Negatives; TP = True Positive; TN = True Negatives.
*Workflow with 100% sensitivity.

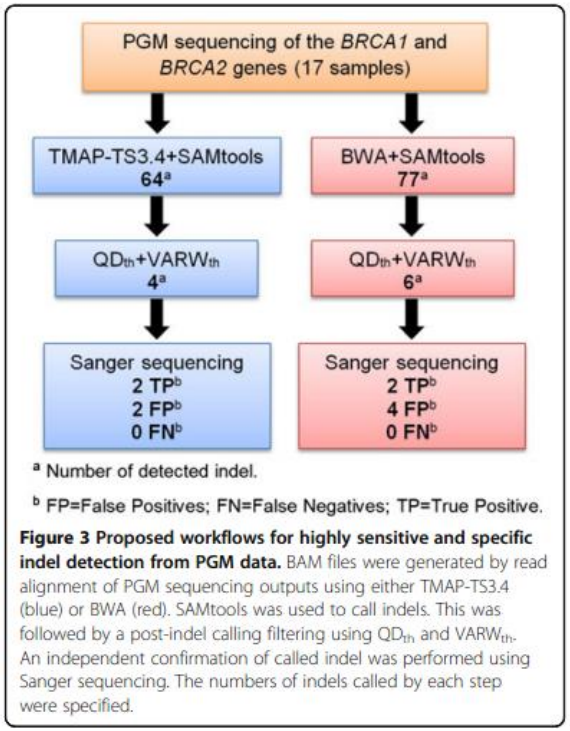


Figure 3 Proposed workflows for highly sensitive and specific indel detection from PGM data. BAM files were generated by read alignment of PGM sequencing outputs using either TMAP-TS3.4 (blue) or BWA (red). SAMtools was used to call indels. This was followed by a post-indel calling filtering using QD_{th} and $VARW_{th}$. An independent confirmation of called indel was performed using Sanger sequencing. The numbers of indels called by each step were specified.

1. Performance evaluation of the torrent suite for indel detection shows an improvement in sensitivity for version 3.4 as compared to the older versions 2.0/2.2.
2. The three TS variant callers were unable to achieve 100% sensitivity. Both GATK and SAMtools achieved 100% sensitivity and 99% specificity on alignment data generated by TS3.4. GATK also performed better than TSVC when calling indels from alignment data of TS2.2. Along with higher sensitivity, both GATK and SAMtools had a lower specificity than TSVC.
3. GATK and SAMtools were also used to call indels from alignment data generated by BWA. The sensitivity of indel calling using both GATK and SAMtools remained as 100%.

We present a computational workflow that uses (1) uses the TS3.4 or BWA as the read mapper (2) SAMtools as the variant caller and (3) VARWth and QDth as post variant-calling filters. This workflow resulted in indel detection with overall 100% sensitivity , >=99.99% specificity and <=44% FDR of all 23 samples.