Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines

刘博

2021.11.4

方法学概览

临床预期用途和选择依据

- 临床预期用途
- 研究背景
- 适用人群

方法建立与优化

- 检测系统设计
- 样本采集要求
- 湿实验流程建立
- 信息分析流程建立和性能确认
- 测序平台性能确认
- 体细胞突变解读流程建立

分析性能确认

- 阳性判断值的确定
- 精密度测试
- 最低检测限
- 空白检测限
- 准确性测试
- 干扰物质
- 交叉反应
- 临床有效性

全周期质控

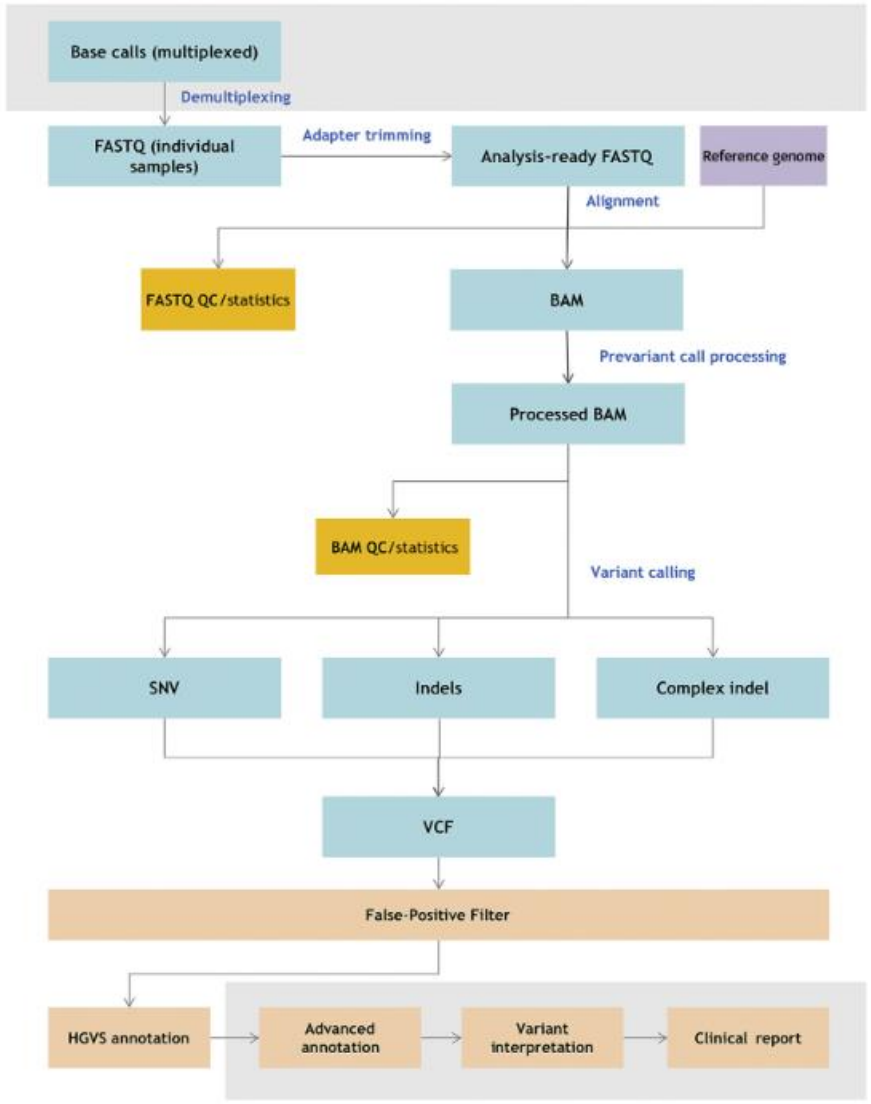
- 要素控制
- 过程控制
- 文档受控

➢ 方法设计参考依据

✓ 李金明，高通量测序技术[M].
✓ 李金明，个体化医疗中的临床分子诊断[M].
✓ Rehm H L, et al. ACMG clinical laboratory standards for next-generation sequencing[J] 2013.
✓ Aziz N, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests[J] 2015.
✓ Matthijs G, et al. Guidelines for diagnostic next-generation sequencing[J] 2016.
✓ Jennings L J, et al. Guidelines for validation of next-generation sequencing–based oncology panels[J] 2017.
✓ Li M M, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer[J] 2017.
✓ Roy S, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines[J] 2018.

✓ 试剂盒性能验证内容相比IVDs无差异
✓ 从实验-生信分析-解读-报告全流程都需要申报，比IVDs只侧重试剂盒性能更全面，更适合复杂的大panel申报

方法建立与优化



CRAM format specification version 3.0; *http://samtools.github.io/hts-specs/CRAMv3.pdf*

VCF; *https://samtools.github.io/hts-specs/VCFv4.3.pdf*

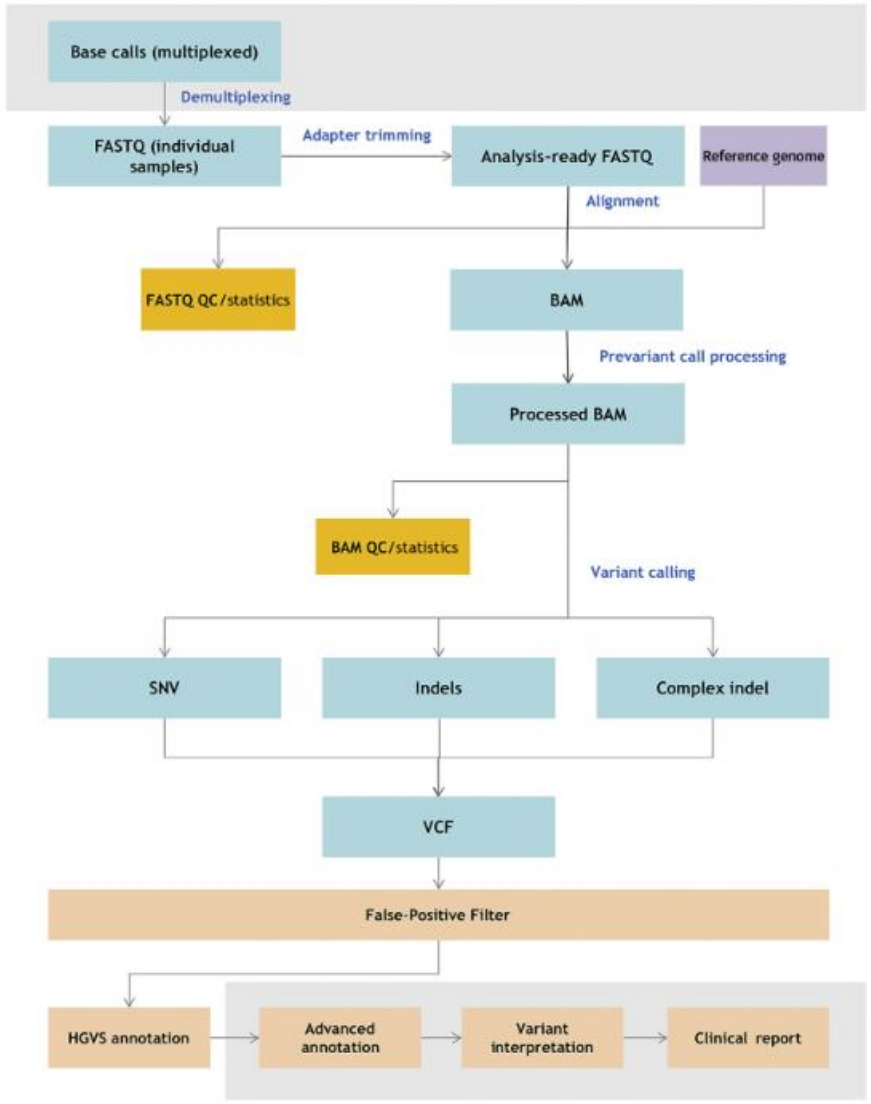
Genomic VCF Conventions, *https://sites.google.com/site/gvcftools/home/about-gvcf/gvcf-conventions*

The Sequence Ontology Genome Variation Format Version 1.10, *https://github.com/The-Sequence-Ontology/Specifications/blob/master/gvf.md*

The Human Genome Variation Society, Human Genome Variation Society (HGVS) Simple Version 15.11. 2016, *http://varnomen.hgvs.org/bg-material/simple*

MAF Mutation Annotation Format；https://docs.gdc.cancer.gov/Encyclopedia/pages/Mutation_Annotation_Format_TCGAv2/

方法建立与优化



AMP Working Group Charge and Scope

This expert working group recommends factors and best practice guidelines for analytical validation of NGS bioinformatics pipelines for detection of SNVs, indels, and multinucleotide substitutions (delins in HGVS terminology) comprising **a length of 21 bp or less** from both somatic and germline human origin (herein referred to as small sequence variants)

Limitations of This Publication

these guidelines do not address the analytical validation of bioinformatics pipelines for large indels >21 bp in length, structural variants (inversions and translocations), gene fusion variants and translocations, gene expression variations, epigenetic variants, copy number alterations, and other variants not defined as SNVs or small indels (herein referred to as large variants). Bioinformatics pipelines designed to detect large variants may be different and less common than general purpose, small sequence variant calling algorithms.

Recommendation

Recommendation 1: Clinical Laboratories Offering NGS-Based Testing Should Perform Their Own Validation of the Bioinformatics Pipeline

Recommendation 2: A Qualified Medical Professional with Appropriate Training in NGS Interpretation and Certification Must Oversee and Be Involved in the Validation Process

Recommendation 3: Validation Must Be Performed Only after Completion of Design, Development, Optimization, and Familiarization of the Bioinformatics Pipeline and Its Components

Recommendation 4: Bioinformatics Pipeline Validation Should Closely Emulate the Real-World Environment of the Laboratory in which the Test Is Performed

Recommendation 5: Validation Should Include All Individual Components of the Bioinformatics Pipeline Used in the Analysis, and Each Component Must Be Reviewed and Approved by an Appropriately Qualified Medical Molecular Professional and the Laboratory Director

Recommendation 6: The Design and Implementation of the Bioinformatics Pipeline Must Ensure the Security of Identifiable Patient Information and Be Compliant with All Applicable Laws at the Local, State, and National Levels

Recommendation 7: Validation of the NGS Bioinformatics Pipeline Must Be Appropriate and Applicable for the Intended Clinical Use, Specimen, and Variant Types Detected by the NGS Test

Recommendation 8: Laboratories Must Ensure That the Design, Implementation, and Validation of the Bioinformatics Pipeline Are Compliant with Applicable Laboratory Accreditation Standards and Regulations

Recommendation 9: The Bioinformatics Pipeline Is Part of the Test Procedure, and Its Components and Processes Must Be Documented according to Laboratory Accreditation Standards and Regulations

Recommendation 10: The Identity of the Sample Must Be Preserved throughout Each Step of the NGS Bioinformatics Pipeline with a Minimum of Four Unique Identifiers, Including a Unique Location Identifier within the Content of Each Data File Read and/or Generated by the Pipeline

Recommendation 11: Specific Quality Control and Quality Assurance Parameters Must Be Evaluated during Validation and Used to Determine Satisfactory Performance of the Bioinformatics Pipeline

Recommendation 12: The Methods Used to Alter or Filter Sequence Reads at Any Point in the Bioinformatics Pipeline before Interpretation Must Be Validated to Ensure That the Data Presented for Interpretation Accurately and Reproducibly Represent the Sequence in the Specimen, and Full Documentation of These Methods Must Be Kept as Part of the Test Documentation according to Laboratory Accreditation Standards and Regulations

Recommendation 13: Laboratories Must Include Specific Measures to Ensure That Each Data File Generated in the Bioinformatics Pipeline Maintains Its Integrity and Provides Alerts for or Prevents the Use of Data Files that Have Been Altered in an Unauthorized or Unintended Manner

Recommendation 14: In Silico Validation Can Be Used to Supplement the Validation of the Bioinformatics Pipeline but Shall Not Be Used in Lieu of End-to-End Validation of Bioinformatics Pipelines Using Human Samples

Recommendation 15: Validation of the Bioinformatics Pipeline Must Include Confirmation of a Representative Set of Variants with High-Quality Independent Data; Appropriate Validation Metrics by Variant Type Should Be Reported

Recommendation 16: Clinical Laboratories Must Ensure the Accuracy of Software-Generated HGVS Variant Nomenclature and Annotations and Have an Alert in Place to Indicate When the Software-Generated Nomenclature or Annotations Need to Be Manually Reviewed and/or Corrected, and Documentation of Any Corrections Must Be Maintained

Recommendation 17: Supplemental Validation Is Required whenever a Significant Change Is Made to Any Component of the Bioinformatics Pipeline

Recommendation

建议 1：提供基于 NGS 的测试的临床实验室应自行验证生物信息学管道

建议 2：在 NGS 解释和认证方面接受过适当培训的合格医疗专业人员必须监督并参与验证过程

建议 3：必须在完成生物信息学管道及其组件的设计、开发、优化和熟悉后进行验证

建议 4：生物信息学管道验证应密切模拟进行测试的实验室的真实环境

建议 5：验证应包括分析中使用的生物信息学管道的所有单个组件，并且每个组件都必须由具有适当资格的医学分子专业人员和实验室主任审查和批准

建议 6：生物信息学管道的设计和实施必须确保可识别患者信息的安全，并遵守地方、州和国家各级的所有适用法律

建议 7：NGS 生物信息学管道的验证必须适合并适用于 NGS 测试检测到的预期临床用途、样本和变异类型

建议 8：实验室必须确保生物信息学管道的设计、实施和验证符合适用的实验室认证标准和法规

建议 9：生物信息学管道是测试程序的一部分，其组成部分和过程必须根据实验室认可标准和法规进行记录

建议 10：必须在 NGS 生物信息学管道的每个步骤中保留样本的身份，至少有四个唯一标识符，包括管道读取和/或生成的每个数据文件内容中的唯一位置标识符

建议 11：在验证期间必须评估特定的质量控制和质量保证参数，并用于确定生物信息学管道的令人满意的性能

建议 12：必须验证用于在解释前在生物信息学管道中的任何点更改或过滤序列读数的方法，以确保提供的用于解释的数据准确且可重复地代表样本中的序列，并且必须提供这些方法的完整文档根据实验室认可标准和法规，作为测试文件的一部分保存

建议 13：实验室必须包括具体措施，以确保在生物信息学管道中生成的每个数据文件保持其完整性，并为以未经授权或无意的方式更改的数据文件的使用提供警报或防止使用

建议 14：计算机验证可用于补充生物信息学管道的验证，但不得代替使用人体样本的生物信息学管道的端到端验证

建议 15：生物信息学管道的验证必须包括对具有高质量独立数据的一组代表性变体的确认；应按变体类型报告适当的验证指标

建议 16：临床实验室必须确保软件生成的 HGVS 变体命名法和注释的准确性，并在需要手动审查和/或更正软件生成的命名法或注释时发出警报，并且必须记录任何更正被维护

建议 17：每当对生物信息学管道的任何组件进行重大更改时，都需要进行补充验证

Recommendation 7

Recommendation 7: Validation of the NGS Bioinformatics Pipeline Must Be Appropriate and Applicable for the Intended Clinical Use, Specimen, and Variant Types Detected by the NGS Test

LOD and Variant Allele Fraction Reference Ranges

For an NGS bioinformatics pipeline, the LOD is represented by two data points: **the minimum required depth of coverage at the variant site** and **the minimum variant allele fraction**, both of whichhavetobe satisfied before a variant can be positively called.

LoD测试样本，要尽可能**贴近预期的深度和预期的浓度**。如果缺乏适合的样本，可以对高频样本进行稀释获得合适的样本。

Contiguous Genetic Regions

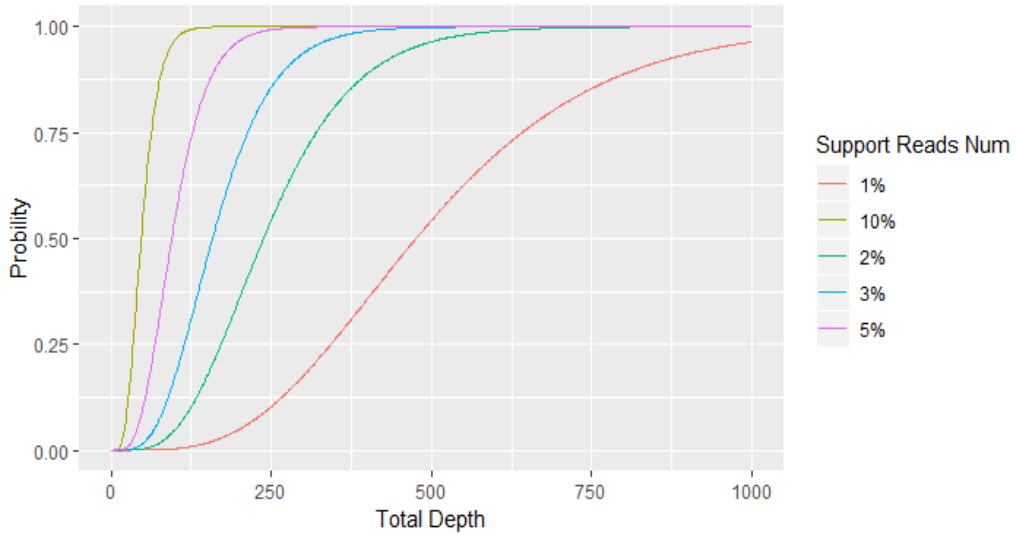
NGS 检测感兴趣区域内的连续遗传区域可能具有不同的序列背景（例如，低复杂性、富含 GC 和同聚序列）。这种序列特征对这些遗传区域中的变异进行充分测序和检测提出了挑战。因此，NGS 检测中包含的每个连续遗传区域的序列质量应在整个验证样本队列中进行分析，以识别测序不良的区域。在分析连续遗传区域时，应包括质量指标的分布，例如覆盖深度、碱基质量、作图质量和链偏差。还应包括代表测定中不同遗传区域的变体，包括测序不良的区域，以确保管道可以准确检测变体或确认管道限制。例如，基因组中富含 GC 的区域（例如，TERT 启动子和 CEBPA）天生就难以测序，并且经常导致低覆盖度、较低的碱基和作图质量以及高链偏倚，最终会影响变异识别的灵敏度。
如果用于给定测序策略的生物信息学算法无法达到这些区域中变异调用的分析灵敏度和阳性预测值所需的水平，实验室必须明确说明所提供的 NGS 测试的局限性，如果需要，选择验证一个解决变异检测的替代测序方法。

原来深度确认的方法

模型设计：预期达到1%检测下限

整个测序检测过程是随机采样的；基于经验先确定可信的条件为：

5条Reads & 有不同的正负链支持



	20%	10%	5%	3%	2%	1%
90%	40	81	163	273	410	822
95%	45	93	187	314	471	945
99%	58	119	240	403	606	1215
99.9%	76	156	316	529	796	1596

不同频率不同检出率的深度要求

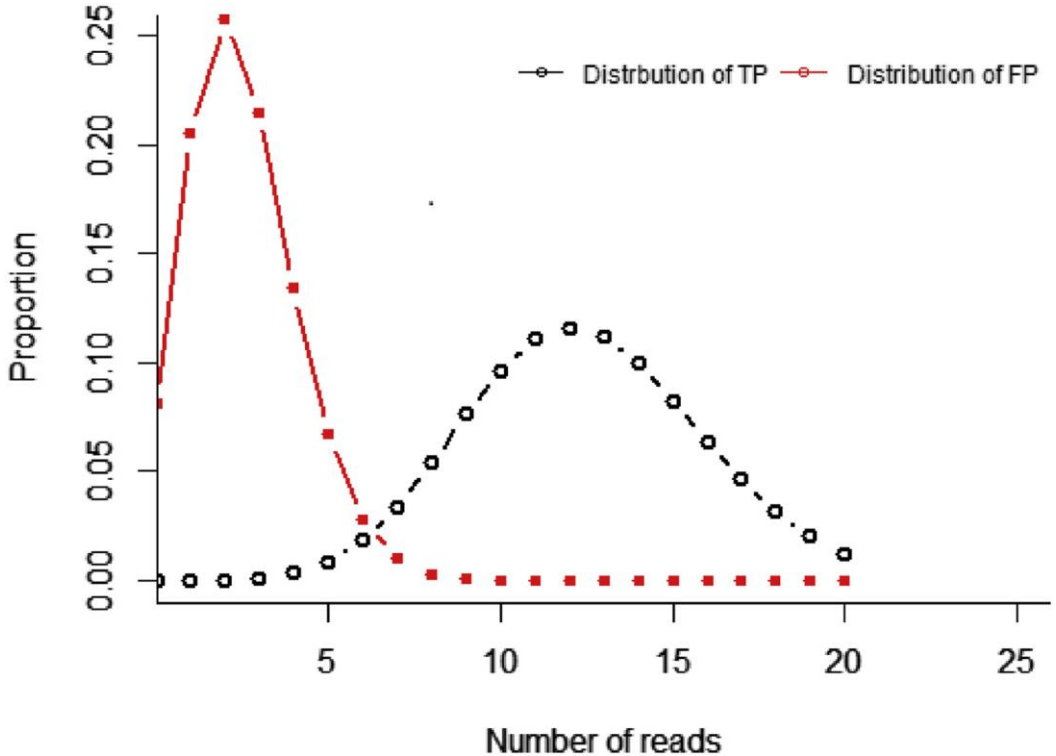
Establishing Criteria for Depth of Sequencing

$$P(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

P(x) is the probability of x variant reads,
x is the number of variant reads,
n is the number of total reads,
p is the probability of detecting a variant allele (ie, the proportion of mutant alleles in the sample).

随机采样的过程，频率确定的情况下，可以计算特定深度下的Reads的分布密度曲线。
通过检测限和错误率，构建各自的分布，确定测序深度的标准

Determining depth of sequence. Given an allele burden of 5% and 250 read depth, the binomial distribution of true positives (TPs) can be calculated. Also, given a sequence error rate of 1%, the binomial distribution of false-positive (FP) results can also be calculated and shown to overlap the true positive distribution. The overlap of true-positive and false-positive distributions should be considered when determining minimum depth of sequence needed to reliably detect a given allele burden

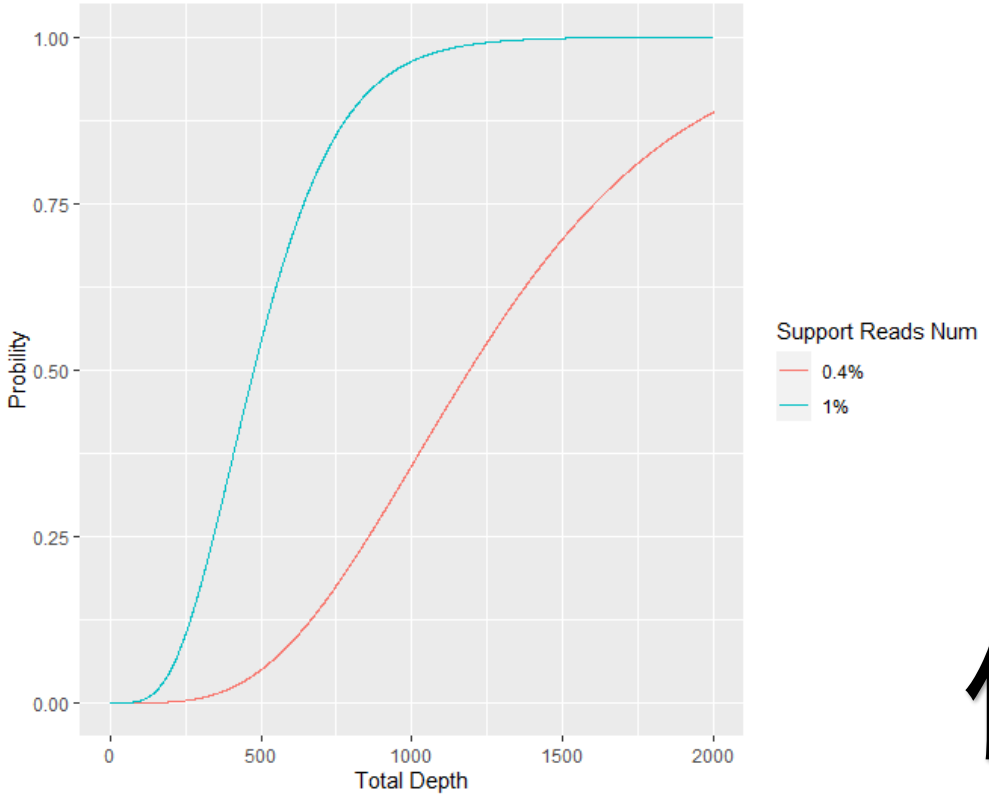


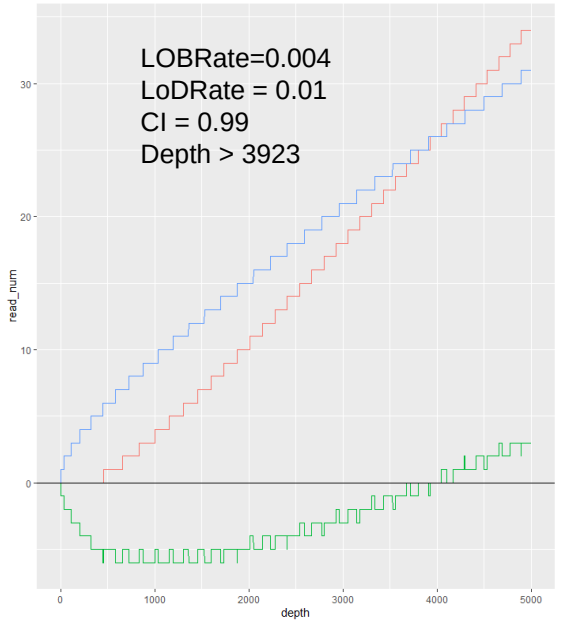
Establishing Criteria for Depth of Sequencing

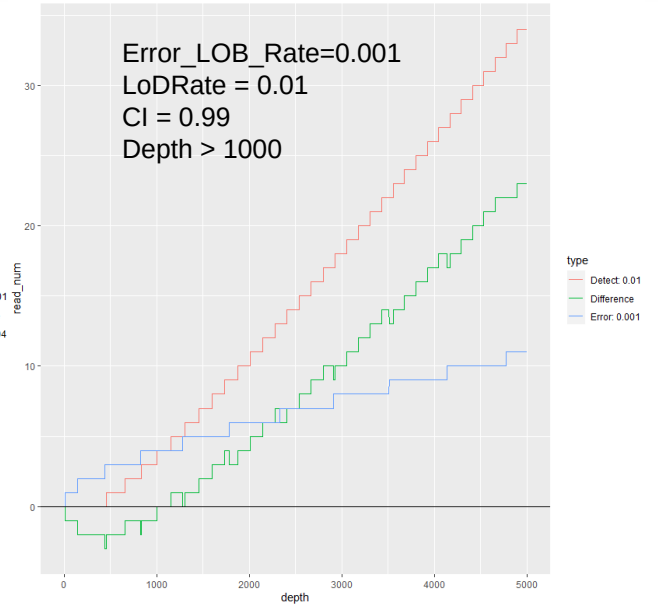
以泛癌历史LDT数据为例进行展示。

	20%	10%	5%	3%	2%	1%
90%	40	81	163	273	410	822
95%	45	93	187	314	471	945
99%	58	119	240	403	606	1215
99.9%	76	156	316	529	796	1596

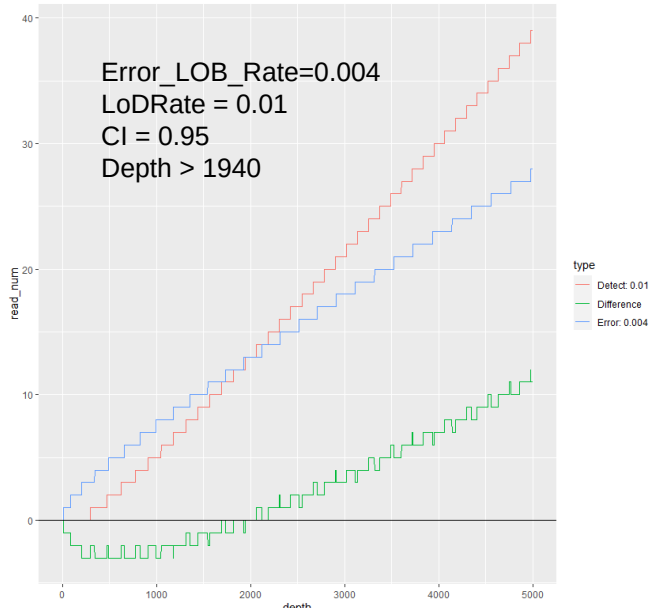
不同频率不同检出率的深度要求







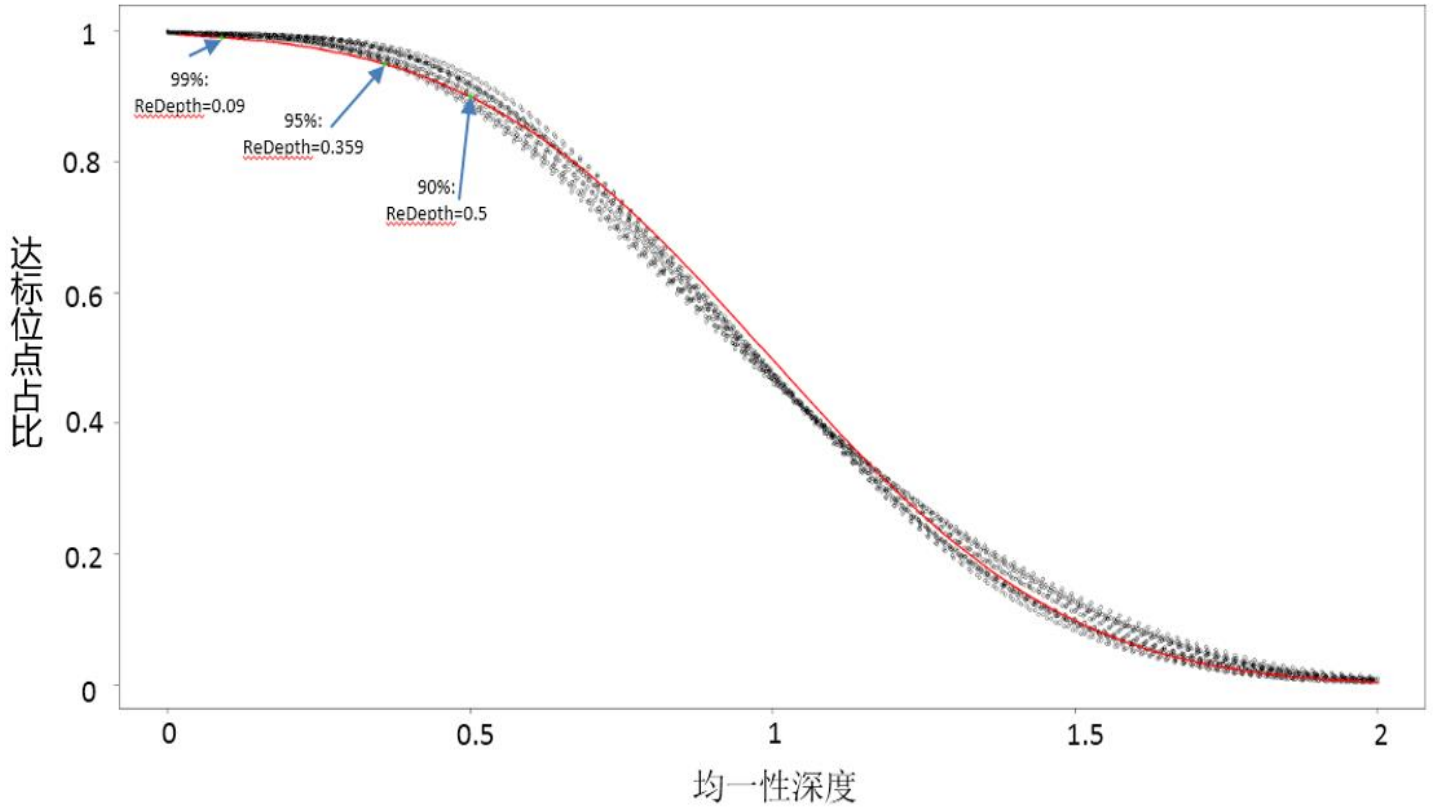
仅当系统错误率达到1‰以下时；
深度评估才有效。
系统错误率过高时，
错误影响便不可忽视。



位点深度！

位点深度 → 样本平均深度

Panel整体覆盖度



位点深度；
样本平均深度；
100x、500x、*x覆盖度

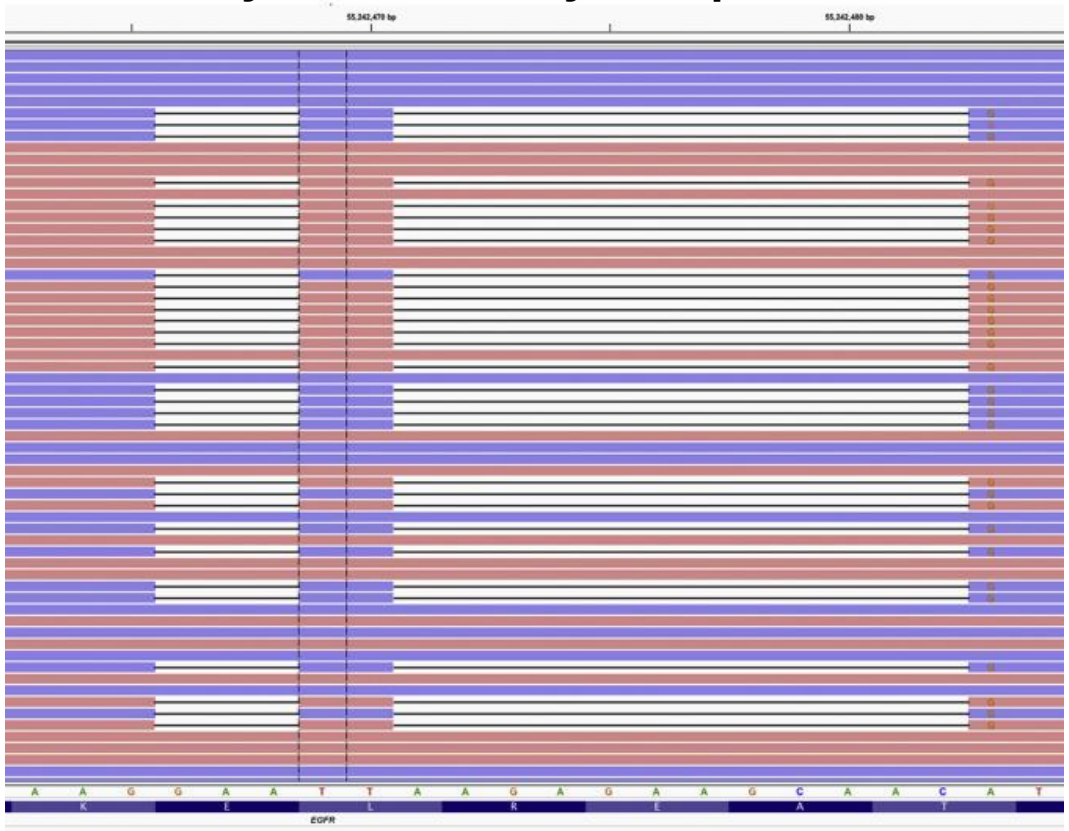
结合位点深度和芯片均一性；制定让更多区域有效的质控深度

471/0.359≈*1300x*

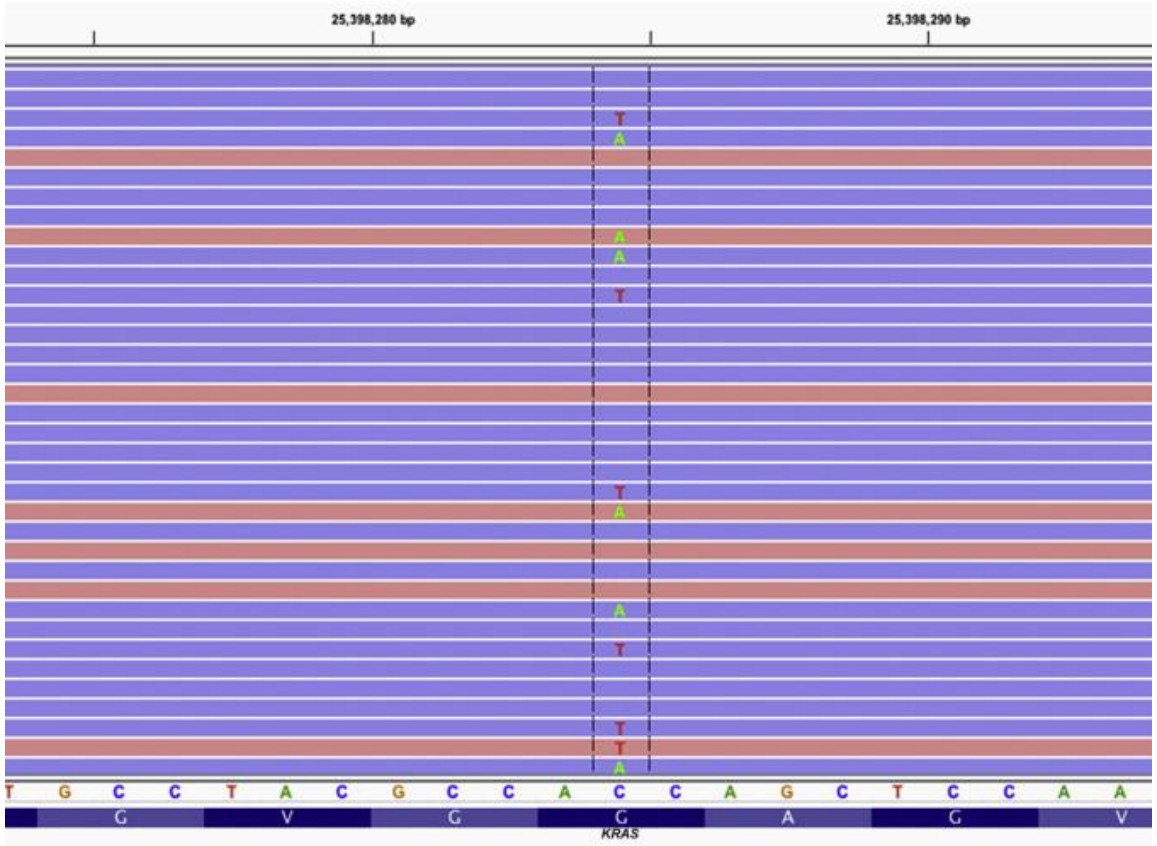
Recommendation 7

Recommendation 7: Validation of the NGS Bioinformatics Pipeline Must Be Appropriate and Applicable for the Intended Clinical Use, Specimen, and Variant Types Detected by the NGS Test

Horizontally and Vertically Complex Variants



An example of a **horizontally** complex variant in exon 19 of the *EGFR* gene



An example of a **vertically** complex variant in exon 2 of the *KRAS* gene

Variants that Require Additional Algorithms

由于序列变化的复杂性或基因组中出现特定变异类型的区域的复杂性，使用通用变异检测算法固有地难以检测某些变异类型。 特定算法通常旨在提高检测此类变体的灵敏度。例如，FLT3 内部串联重复是急性髓系白血病的临床显着变异，需要使用常规变异调用者之外的特定算法进行检测。

Recommendation 7

Recommendation 7: Validation of the NGS Bioinformatics Pipeline Must Be Appropriate and Applicable for the Intended Clinical Use, Specimen, and Variant Types Detected by the NGS Test

样本数：Minimum Number of Wet Laboratory Samples to Include in the Validation Sample Set

$$\sum_{i=k}^{n}\binom{n}{i}p^{n-i}(1-p)^{i} = 1 - CL \quad (2)$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (3)$$

when k is an integer between 0 and n, 0 k n and CL is the confidence level (eg, 0.95). By setting $k = 0$ (ie, 0 failures), the formula can be simplified to: $p^n = 1 - CL$

$$n = \frac{\ln(1-CL)}{\ln(p)} \quad (4)$$

基于需要的灵敏性和置信区间，确定样本数

$$n = {\ln(1-0.95)}/{\ln(0.95)}$$

$$\approx {-2.996}/{-0.051}$$

$$\approx 58.4 \approx 59$$

变异数：Minimum Number of Variants to Include in the Validation Sample Set

This is the same calculation that is used to determine the number of wet laboratory samples in valida- tion of an NGS analysis for cancer : $n = {\ln(1-\mathrm{CL})}/{\ln(\mathrm{p})}$, where n is the number the confidence level of detection, and P is the probability of detection.

SNV、InDel、复杂变异等，每个细分类型的变异，测试数目达到59个。

Recommendation 7

Recommendation 7: Validation of the NGS Bioinformatics Pipeline Must Be Appropriate and Applicable for the Intended Clinical Use, Specimen, and Variant Types Detected by the NGS Test

Failure to Detect a Variant in the Validation Sample Set

如果发现任何一个变异失败的时候，要详细的核查检测失败的原因。检查Bam文件就是一个非常必要的核查方案，比如使用IGV（ Integrated Genomics Viewer ）。如果核查确定是样本被破坏或和临床常规检测样本不一致，则可以进行样本进行替换。但如果样本适合本次检测，则需要对流程进行调查，要么对流程部分组件进行调整，要么在临床报告中说明此类限制。

Recommendation 10

Recommendation 10: The Identity of the Sample Must Be Preserved throughout Each Step of the NGS Bioinformatics Pipeline with a Minimum of Four Unique Identifiers, Including a Unique Location Identifier within the Content of Each Data File Read and/or Generated by the Pipeline

在NGS流程生成设计的所有数据结果中，都应该标记下列四个标记符号，用来区分样本
(FASTQ, sequence alignment/map/BAM, and VCF or equivalent):
i) a unique sample identifier,
ii) a unique patient identifier,
iii) a unique run identifier,
iv) a laboratory location identifie

如果实验室有一个绝对唯一的代码可以区分样本、患者、实验室，那可以去掉这三个部分；
但是分析代码必须是保留。

http://www.hl7.org/

Recommendation 11

Recommendation 11: Specific Quality Control and Quality Assurance Parameters Must Be Evaluated during Validation and Used to Determine Satisfactory Performance of the Bioinformatics Pipeline

Table 4 Recommended Quality Metrics for Clinical Bioinformatics Pipelines

Category	Use	Quality metric	Performance criteria*	Used for
Preanalytical	REQ	% of nucleated cells that are tumor cells	Min	Tumor samples
Sample	REQ	DNA concentration	Min, Max	All sample types
Sample	REQ	DNA fragment size	Min, Max	All sample types
Sample	REQ	Library DNA quantification	Min	All sample types
Run metrics	REQ	Cluster density	Min, Max	All sample types on Illumina platforms that include this metric by default
Run metrics	REQ	% of bases higher than the minimum Phred score of all bases called	Min	All sample types
Run metrics	REQ	Demultiplexing success (ie, all molecular identifiers present and no unexpected molecular identifiers detected)	Pass/fail	All sample types when multiplexing is used
Run metrics	REQ	% of reads passing a minimum Phred score criterion (eg, 99% of bases at Q30 or higher)	Min	All sample types
Read filters	REQ	Mapping quality	Min	All sample types
Mapping†	REQ	Mean on-target coverage of reads	Min	All sample types
Mapping†	REQ	% of targeted bases with coverage greater than a specified minimum	Min	All sample types
Mapping†	REQ	% of bases exceeding the minimum Phred score mapped on target	Min	All sample types
Mapping†	OHR	% of aligned bases exceeding the minimum Phred score that disagree with reference	Max	Samples for germline analysis only
Mapping†	OHR	AT/GC bias	Max	All sample types
Mapping†	REQ	Mean insert size (bp)	Min, Max	All sample types for hybrid capture methods only
Mapping†	REQ	% PCR duplicates	Max	All sample types using non–amplicon-based sequencing
Per variant	REQ	Depth of coverage at variant's position	Min	All sample types
Per variant	REQ	Quality score	Min	All sample types
Per variant	Opt	Number of germline SNVs	Min, Max (may have to have separate criteria for different ethnicities)	All sample types
Per variant	REQ	Allele fraction	Min	All sample types
Per variant	REQ	Strand bias	Max	All sample types
Per variant	Opt	Haplotype bias	Max	All sample types
Per variant	REQ	Number of distinct vertical variants at the same position	≤2	All sample types
Per variant	REQ	Number of distinct horizontal variants within a prescribed cluster window size (bp)	≤1	All sample types
QC†	OHR	Estimate of % contamination from another sample	Max	Samples for germline analysis only (optional for tumor samples)
QC†	Opt	Fingerprint genotypes match NGS results	Yes (no requires investigation/explanation)	All sample types
QC†	REQ	Observed sex matches reported sex	Yes (no requires investigation/explanation)	All sample types if X/Y chromosomes are included in assay
QC†	Opt	% of bases called that are variants	Min, Max	Samples for germline analysis only (optional for tumor samples)
QC†	Opt	SNP/indel ratio	Min, Max	All sample types
QC†	Opt	Ti/Tv ratio	Min, Max	All sample types
QC†	Opt	Ratio of heterozygous/homozygous variants	Min, Max	Samples for germline testing only
QC†	Opt	Coverage profile compared with controls	Goodness-of-fit test	Critical for copy number analysis but also useful for assay QC

Max，最大阈值（高于该值的样本或指标被视为不可接受或失败）；
Min，最小阈值（低于该值，样本或指标被视为不可接受或失败）；
NGS，新一代测序；
Opt，可选；
QC，质量控制；
SNV，单核苷酸变体；
Indel，插入/删除；

OHR，可选但强烈推荐；
Q30，Phred（质量）得分 Z 30；
REQ，对于指定的样品类型是必需的；
Ti/Tv，转换次数/转换次数

Recommendation 11

Mapping†	OHR	% of aligned bases exceeding the minimum Phred score that disagree with reference	Max	Samples for germline analysis only
Mapping†	OHR	AT/GC bias	Max	All sample types
Mapping†	REQ	Mean insert size (bp)	Min, Max	All sample types for hybrid capture methods only
Mapping†	REQ	% PCR duplicates	Max	All sample types using non–amplicon-based sequencing
Per variant	REQ	Depth of coverage at variant's position	Min	All sample types
Per variant	REQ	Quality score	Min	All sample types
Per variant	Opt	Number of germline SNVs	Min, Max (may have to have separate criteria for different ethnicities)	All sample types
Per variant	REQ	Allele fraction	Min	All sample types
Per variant	REQ	Strand bias	Max	All sample types
Per variant	Opt	Haplotype bias	Max	All sample types
Per variant	REQ	Number of distinct vertical variants at the same position	≤2	All sample types
Per variant	REQ	Number of distinct horizontal variants within a prescribed cluster window size (bp)	≤1	All sample types
QC†	OHR	Estimate of % contamination from another sample	Max	Samples for germline analysis only (optional for tumor samples)
QC†	Opt	Fingerprint genotypes match NGS results	Yes (no requires investigation/ explanation)	All sample types
QC†	Opt	% of bases called that are variants	Min, Max	Samples for germline analysis only (optional for tumor samples)
QC†	Opt	SNP/indel ratio	Min, Max	All sample types
QC†	Opt	Ti/Tv ratio	Min, Max	All sample types

Recommendation 13

Recommendation 13: Laboratories Must Include Specific Measures to Ensure That Each Data File Generated in the Bioinformatics Pipeline Maintains Its Integrity and Provides Alerts for or Prevents the Use of Data Files that Have Been Altered in an Unauthorized or Unintended Manner

生信分析结果产生的大文件，在进行数据转移、上传、下载等过程可能因为硬件问题等原因导致文件异常。
实验室要使用 a hash/checksum method 来校验确保文件完整性。

MIT Laboratory for Computer - Science and RSA Data Security, Inc.,
The MD5 Message- Digest Algorithm, https://www.ietf.org/rfc/rfc1321.txt,last accessed September 26, 2017; or
National Institute of Standards and Technology Computer Security Resource Center, Hash Functions,
https://csrc.nist.gov/Projects/Hash- Functions/publications, last accessed September 26, 2017

Recommendation 14

Recommendation 14: In Silico Validation Can Be Used to Supplement the Validation of the Bioinformatics Pipeline but Shall Not Be Used in Lieu of End-to-End Validation of Bioinformatics Pipelines Using Human Samples

模拟数据可以用来辅助进行测试，但是不能替代真实样本。

Simulator	Technology	G vs M	Run types	Ref seq	Characterization							Processes			Outputs		
					Input			Profile process									
					PA	RE	PR	DF	PA	GU	SW	PCR	GV	QS	RE	AL	FO
454sim	454	G	SE	Yes	No	No	Yes	Yes	No	No	No	No	No	Yes	Yes	No	SFF
ART	454, Illumina and SOLiD	G	SE, PE and MP	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	Yes	SFF and FQ
ArtificialFastqGenerator	Illumina	G	PE	Yes	Yes	Yes	No	No	Yes	No	No	No	No	Yes	Yes	No	FQ
BEAR	454, Illumina and IonTorrent	G and M	SE and PE	Yes	No	Yes	No	No	No	Yes	No	No	Yes	Yes	Yes	No	FQ
CuReSim	454, Illumina, SOLiD and IonTorrent	G	SE	Yes	Yes	No	No	No	Yes	No	No	No	No	Yes	No	No	FQ
DWGSIM (DNA analysis)	Illumina, SOLiD and IonTorrent	G	SE, PE and MP	Yes	Yes	No	Yes	No	Yes	No	No	No	Yes	Yes	Yes	No	FQ
EAGLE	454, Illumina, PacBio and IonTorrent	G	SE and PE	Yes	No	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	FQ
FASTQSim	Illumina, SOLiD, PacBio and IonTorrent	G and M	SE	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	FQ
Flowsim	454	G	SE and PE	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	No	SFF
GemSim	454 and Illumina	G and M	SE and PE	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	FQ
Grinder	454, Illumina and Sanger	G and M	SE, PE and MP	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	No	FQ
Mason	454, Illumina and Sanger	G	SE, PE and MP	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	FA and FQ
MetaSim	454, Illumina and Sanger	G and M	SE, PE and MP	Yes	Yes	No	No	No	Yes	No	No	No	No	No	Yes	No	FA
NeSSM	454 and Illumina	M	SE and PE	Yes	No	No	Yes	No	No	Yes	No	No	Yes	Yes	Yes	No	FQ
pbsim	PacBio	G	CLR and CCS	Yes	Yes	No	No	Yes	No	No	No	No	No	Yes	Yes	Yes	FQ
pIRS	Illumina	G and M	PE	Yes	Yes	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	No	FQ
ReadSim	PacBio and Nanopore	G	SE	Yes	Yes	No	No	No	Yes	No	No	No	Yes	Yes	Yes	No	FQ
simhtsd	454 and Illumina	G	SE and PE	Yes	Yes	No	No	No	Yes	No	No	No	No	Yes	No	No	FQ
simNGS	Illumina	G	SE and PE	Yes	Yes	No	Yes	Yes	No	No	No	No	No	Yes	Yes	No	FQ
SimSeq	Illumina	G	SE, PE and MP	Yes	Yes	No	Yes	Yes	No	No	Yes	No	Yes	Yes	No	Yes	SAM and BAM[#]
SInC	Illumina	G	PE	Yes	No	Yes	Yes	No	No	No	Yes	No	Yes	Yes	Yes	No	FQ
wgsim	Illumina and SOLiD	G	SE	Yes	Yes	No	No	No	Yes	No	No	No	Yes	Yes	Yes	No	FQ
XS	454, Illumina, SOLiD and IonTorrent	G	SE and PE	No	Yes	No	No	No	Yes	No	No	No	No	Yes	Yes	No	FQ

Simulators	Genomic variants*							
	MGC	PLO	SNPs	Indels	INVs	TRA	CNVs	STRs
BEAR	Yes	No	No	No	No	No	No	No
DWGSIM (DNA analysis)	No	Yes	Yes	Yes	Yes	Yes	No	No
EAGLE	No	Yes	Yes	Yes	Yes	Yes	Yes	No
FASTQSim	No	No	Yes	Yes	No	No	No	Yes
GemSim	Yes	No	Yes	Yes	No	No	No	No
Grinder	Yes	No	Yes	Yes	No	No	No	No
Mason	No	No	Yes	Yes	No	No	No	No
NeSSM	Yes	No	No	No	No	No	No	No
pIRS	No	Yes	Yes	Yes	Yes	No	No	No
ReadSim	No	Yes	Yes	Yes	Yes	No	No	No
SimSeq	No	Yes	No	No	No	No	No	No
SInC	No	No	Yes	Yes	No	No	Yes	No
wgsim	No	Yes	Yes	Yes	No	No	No	No

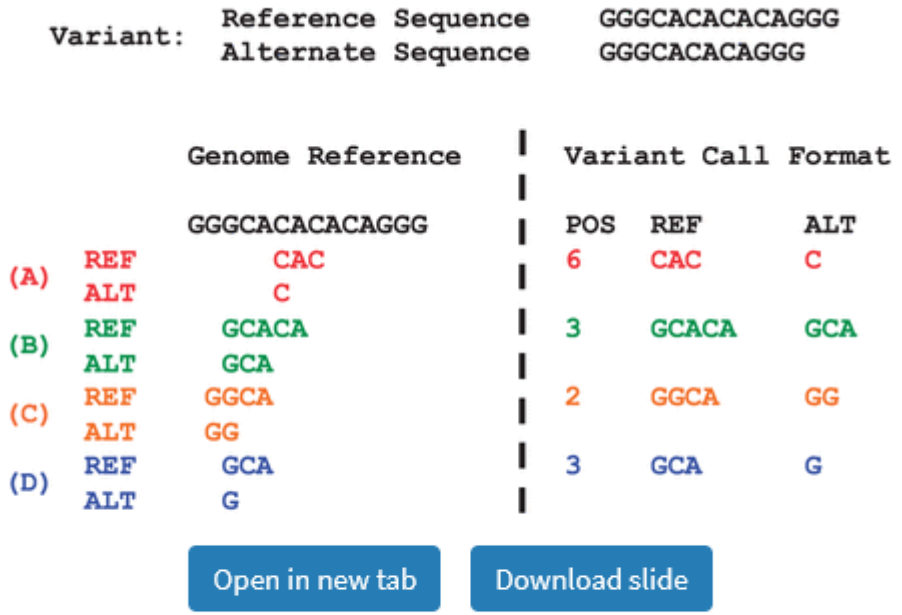
Escalona, M., Rocha, S., & Posada, D. (2016). *A comparison of tools for the simulation of genomic next-generation sequencing data. Nature Reviews Genetics, 17(8), 459–469.* doi:10.1038/nrg.2016.57

Recommendation 16

Recommendation 16: Clinical Laboratories Must Ensure the Accuracy of Software-Generated HGVS Variant Nomenclature and Annotations and Have an Alert in Place to Indicate When the Software-Generated Nomenclature or Annotations Need to Be Manually Reviewed and/or Corrected, and Documentation of Any Corrections Must Be Maintained

Anormalized variant representation in a VCFfile requires that it be parsimonious and left aligned.

Fig. 1.



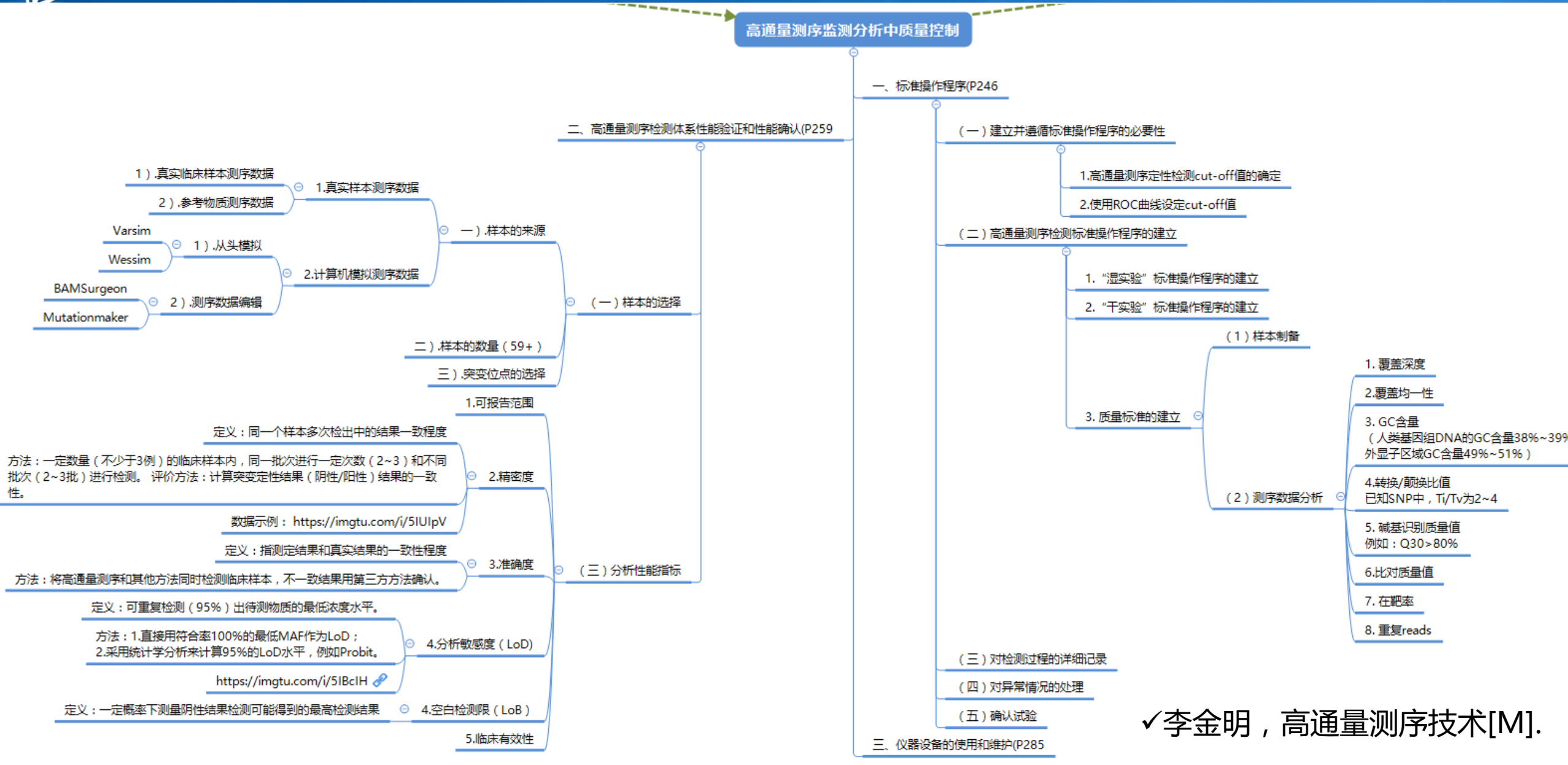
Example of VCF entries representing the same variant. Left panel aligns each allele to the reference genome, and the right panel represents the variant in VCF. (**A**) is not left-aligned (**B**) is neither left-aligned nor parsimonious, (**C**) is not parsimonious and (**D**) is normalized

变异在vcf中是左对齐的，在HGVS中是右对齐(3')。

也提及转录本选择，需要形成文档和记录；但未提供具体方案。

Adrian Tan, Gonçalo R. Abecasis, Hyun Min Kang, Unified representation of genetic variants, *Bioinformatics*, Volume 31, Issue 13, 1 July 2015, Pages 2202–2204, https://doi.org/10.1093/bioinformatics/btv112

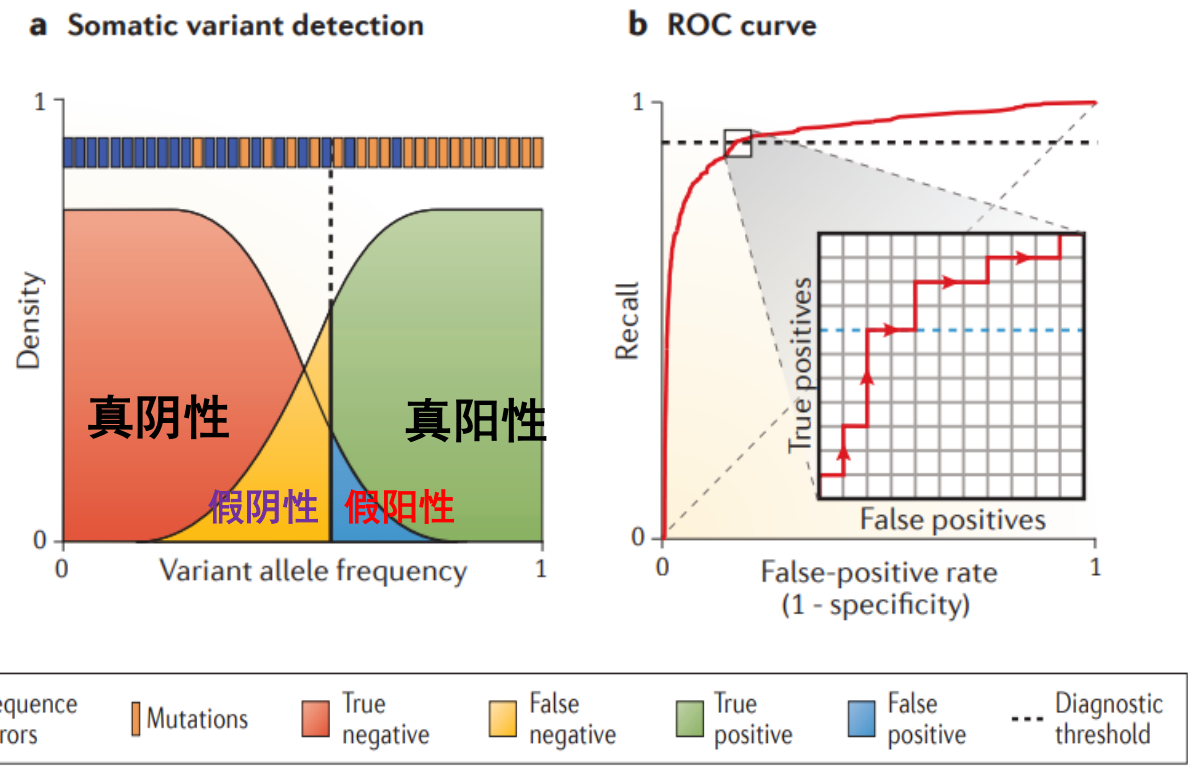
高通量测序监测分析中质量控制



✓李金明，高通量测序技术[M].

阳性判断值的确定

➢ 评估方法的依据

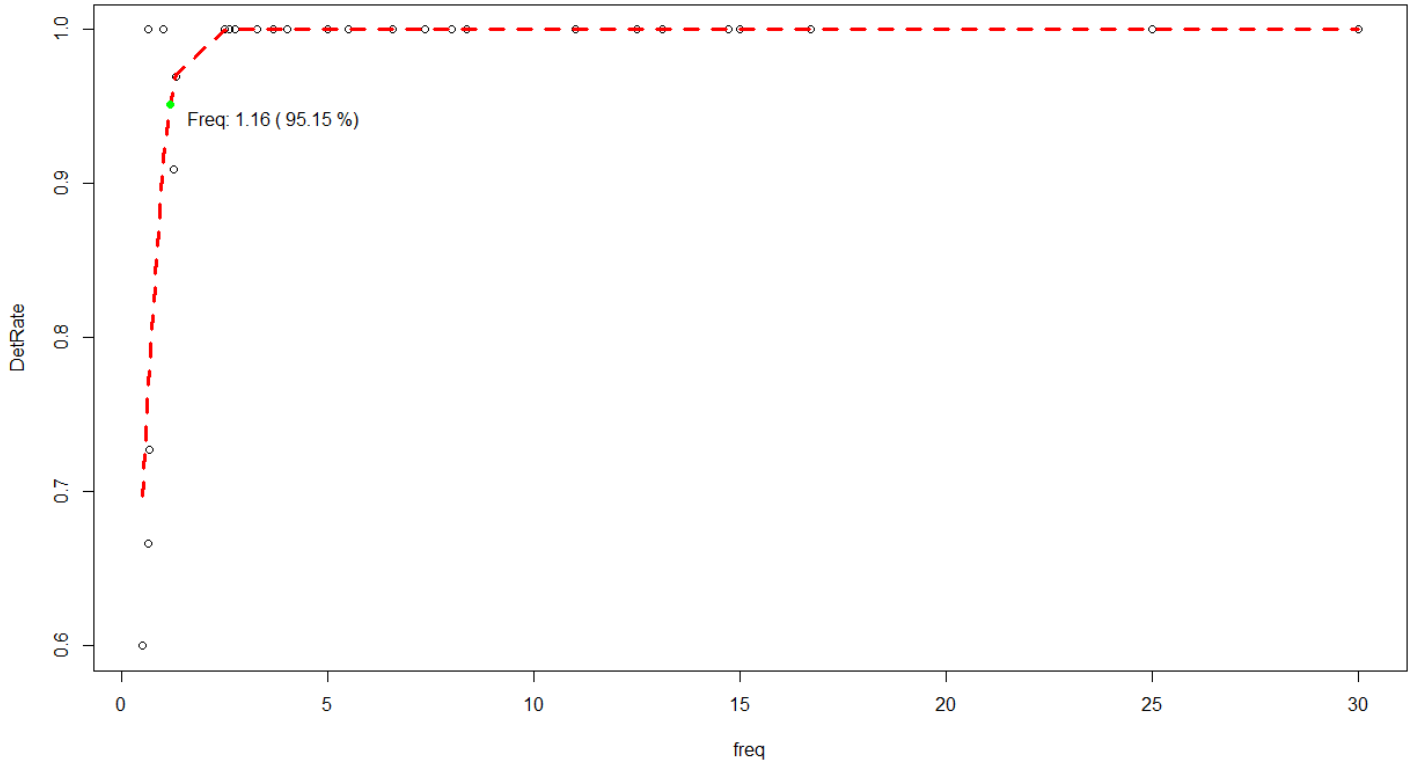


➢ 评估策略：

- 四种变异类型基因分别进行**单个位点的评估**；
- 每个位点分别包含10个阳性和10个阴性临床样本，通过ROC曲线分析得到**每个位点的cut-off值**；
- 针对某一个突变类型确定阈值，最终选取所有位点中最高阈值作为该突变类型的阈值。

生信方法学的建立 - LoD

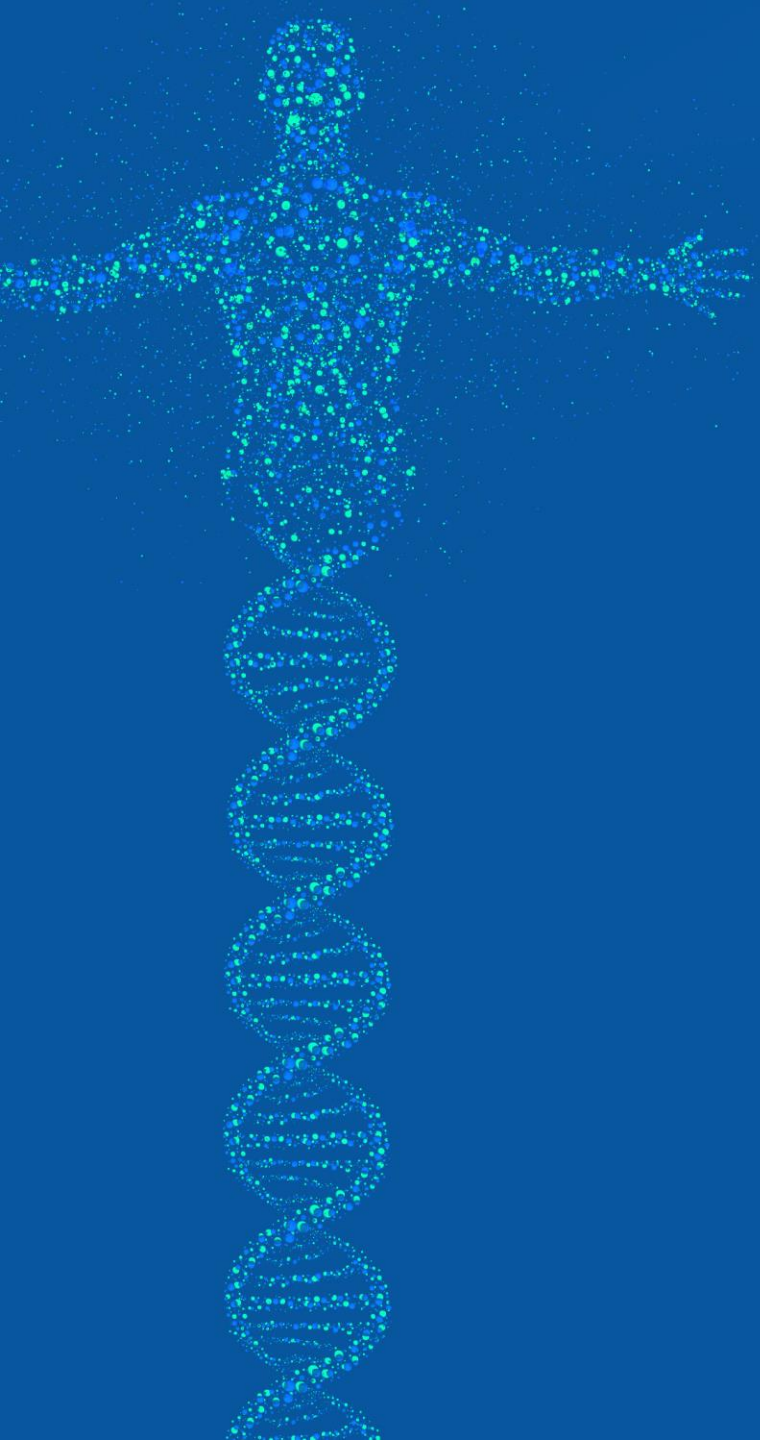
Probit



Probit分析说明：对NGS检测进行简化，认为检出率是仅依赖于变异频率的概率分布函数F(x)；针对每个特定的频率，检出率均满足一个正态分布f(x)。通过测试一定数量的不同频率的变异，和对应的检出率，进行回归拟合获得F(x)的函数曲线。以拟合曲线上检出率为95%时对应的频率作为LOD。

100%

性能参数	样本	重复次数	突变类型	总位点数	评估指标	结论
精密度	5个标准品+16个临床FFPE	批内3次 批间3次	SNV, InDel, CNV, SV	725	阳性符合率	重复性和重现性均为100%
检测下限LoD	10个梯度频率稀释标准品	10次	SNV, InDel, CNV, SV	107	95%稳定检出的MAF	SNV: 2%, InDel: 1.5%, CNV: 4 copies, SV: 1%
空白限LoB	10个临床FFPE	3次	SNV, InDel, CNV, SV	1920	阴性位点检出最高MAF	SNV: 0.3%, InDel: 0.1%, CNV: 2.8 copies, SV: 0.2%
准确性	266个临床样本	1次	SNV, InDel, CNV, SV	1918	与对比方法（Sanger, Arms-PCR, QPCR, ddPCR, NGS）比较变异检出一致性	与临床常用方法比较，准确率为99.77%; 与其他NGS方法对比，阳性符合率为96.25%;
MSI	137个临床样本	1次	/	/	与PCR方法比较MSI检出一致性	与临床方法比较，MSI准确性为91.97%
TMB	准确性76个临床样本； 精密度40个临床样本；	准确性1次； 批内3次+批间3次；	SNV, InDel	/	与WES比较，Panel检测TMB一致性	与WES方法相比，相关系数R^2为0.9496； TMB重复性和重现性%CV均小于20%；



THANK YOU

感谢观看