Detecting Low Frequency SNVs with NGS Sequencing – Introducing VarPROWL

Chad C. Brown[1], Gunjan D. Hariani[1], Matthew C. Schu[1], Keith A. Peoples[1], Rao V. N. Kakuturu[2], Zhancheng Zhang[1], Karen J. Pry[2], Diarmuid M. Moran[2], Sarah Bacus[2], Victor J. Weigman[1]

1) Expression Analysis a Quintiles Company, Durham, NC
2) Quintiles Translational R&D Oncology, Westmont, IL, USA

Introduction

Importance of detecting LFSNVS

Low frequency single nucleotide variants (LFSNVs) have many important medical uses:

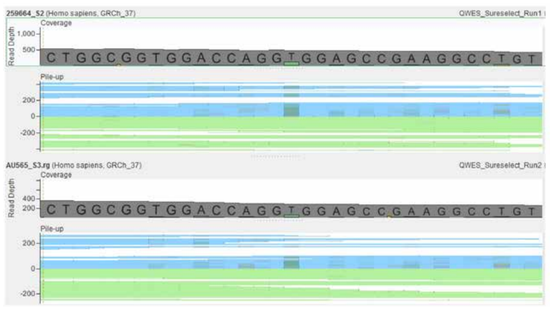
- Deciding which medication is best
- Discovery of resistant subtypes
- Identifying tissue of origin
- Immunology screening

Difficulties in LFSNV detection

- Artifacts related to formalin-fixed, paraffin embedded (FFPE) samples
- Polymerase chain reaction (PCR) artifacts
- Alignment issues (e.g. indels appearing as SNVs)
- Context specific errors (CSEs)
- Low complexity regions
- Improper/ imperfect reference

Context-specific errors

Genomic context causes inflated error rates in genomic sequencing[1, 2]

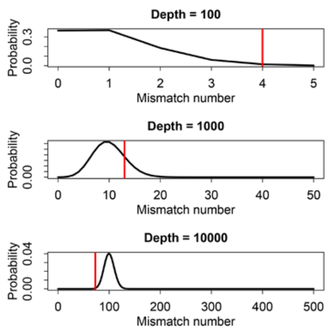


Low complexity regions

Low complexity regions are a special kind of CSE that increases sequencing error rates

The curse of high depth

High depth can increase false positive rates[3], with inaccurate error modeling contributing (e.g. true error = 1% and estimated = 0.5%).



Conclusions

Concluding remarks

VarPROWL is a great tool for detecting LFSNVs

- Direct modeling of sequencer error rates reduces FPs, while not sacrificing sensitivity
- Genomic context (e.g. CSEs) and low complexity are accounted for
- Robust to a wide range of sequencing depths
- Sequencing platform independent (Illumina, PacBio, Ion Torrent)

Future work

- Detection of other mutation types, including indels and CNVs
- Modeling of paired tumor / normal samples
- Porting from R to Java or C
- Incorporation of other data, including mapping quality and PhiX error rates

References

[1] Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer, and Lior Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics*, 12(1):451, 2011.

[2] Manuel Allhoff, Alexander Sch¨onhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC bioinformatics*, 14(Suppl 5):S1, 2013.

[3] Heng Li. Towards better understanding of artifacts in variant calling from high-coverage samples. *arXiv preprint arXiv*:1404.0929, 2014.

[4] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[5] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[6] A Rimmer, I Mathieson, G Lunter, and G McVean. Platypus: an integrated variant caller, 2012.

Data & software

Samples

- Seven cancer samples (FFPE cell lines and FFPE tissue), three non-cancer samples, across four different cancer types
- Sequenced using Quintiles Comprehensive Cancer Panel (QCCP)
- Variants validated using the Ion AmpliSeq™ Cancer Hotspot Panel V2 (ASCP)

ASCP and QCCP

QCCP (Illumina)

- SNVs, indels, copy number variation, and microsatellite instability assessed in coding sequences of 208 cancer, DNA repair, and pharmacogenomic genes
- 17 genes assessed for characterized genomic rearrangements
- Hybridization based enrichment (Agilent)

ASCP (Ion Torrent)

- Assesses hotspots (2800 COSMIC mutations) from 50 genes (22kbp of captured regions)
- PCR based
- PGM NGS platform

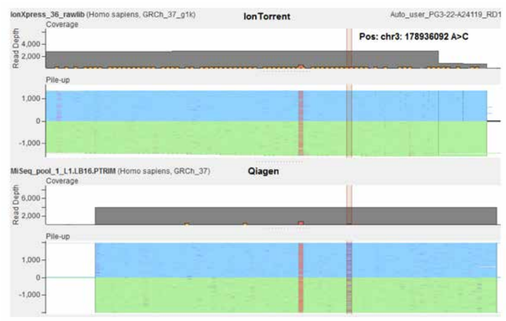
Assay overlap ASCP and QCCP

AmpliSeq™ Hotspot Cancer Panel (v2): 21.8 kb target region

Quintiles Comprehensive Cancer Panel: 1.1mb target reigion

Orthogonal validation

Validation using orthogonal sequencing technologies, mitigates sequencer specific error



Validation procedure

If depth was > 100 for Illumina® (ILL) and > 500 for Ion Torrent (IT), then the following table was used to determine true positive (TP), true negative (TN) or indeterminate (Ind) status. Here, variant (Var) is showing both read directions having allele frequencies > 0:01, otherwise reference (Ref).

Sample	Var on IT Ref on IT	Ref on IT
Var on ILL	TP	Ind
Ref on ILL	Ind	TN

VarPROWL

VarPROWL

- Open source under GNU public license
- Written in R and Perl
- Fast for targeted panels (about 30 minutes for a QCCP sample)
- Allows for parameter tuning to match sequencing chemistry (ILL, IT, etc.), sample type (FF or FFPE) and enrichment type (PCR or hybridization)

Modelling sequencing error rates

Sequencing error rates of forward and reverse reads are estimated independently using a logistic regression model:

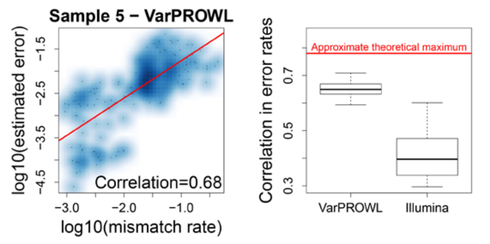
$$\log\left(\frac{e(X)}{1-e(X)}\right) = X\gamma,$$

where:

- X is the regression matrix containing information on complexity metrics, genomic context, nearby error rates, base qualities for reference and non-reference calls and the identity of the non-reference nucleotide.
- $e(X)$ is the sequencing error rate
- γ is estimated using maximum likelihood

Improved error rate modelling

Error rates are modelled more accurately than Illumina's average variant base quality, alone.

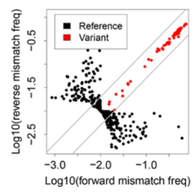


Variation in error rate modelled using beta-binomial

- No model is perfect
- Each read may have different error profile
- Model the distribution (or uncertainty) in error rates using the beta distribution
  e ~ Beta (α,β) *v ~ Binom* (e,d)
- *e* is the sequencing error rate
- *d* is the sequencing depth
- *v* is the non-reference count
- α and β are estimated using maximum likelihood

Bidirectionality check

This filter ensures concordance in variant frequency between read directions.

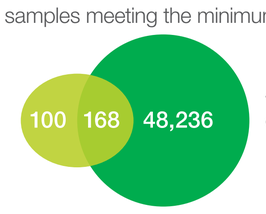


Results

Variant concordance

Variant calls across all samples meeting the minimum QC criteria for each platform

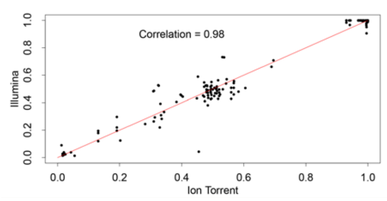
268 variants detected in ASCP



48,404 variants detected in QCCP

Variant linearity

Variant calls across all samples meeting the minimum QC criteria for each platform



Variant calling commands

VarScan 2

java -Xmx1g -jar /opt/downloads/varscan/VarScan.v2.3.6.jar mpileup2snp input.pileup --output-vcf 1 --min-var-freq 0.01 --mincoverage 20 --min-reads2 4 > input.varscan2.liberal.vcf

GATK

java -Xmx2g -jar /opt/downloads/GATK-3.1/GenomeAnalysisTK.jar -T UnifiedGenotyper -R hg19.fa -I input.rg.bam -stand_call_conf 30 -stand_emit_conf 10 -dt NONE -L ../QCCP.annot.bed > input.gatk.vcf

Platypus

python /opt/downloads/Platypus_0.5.2/Platypus.py callVariants --bamFiles=input.rg.bam --refFile= hg19.fa --output=input.platypus.vcf --filterDuplicates=0 --regions=../QCCP.annot.platypus.txt

Comparison variant callers

The features offered for Genome Analysis Toolkit (GATK)[4], VarScan 2 (VS2)[5], Platypus (PP)[6] and VarPROWL are compared. CSE=context specific error and LRA=local realignment/ assembly.

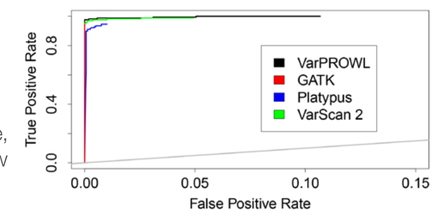
Feature	GATK	VS 2	PP*	VarPROWL
Low complexity			?	X
CSE	X		?	X
Bidirectionality	X	X	X	X
Error modeling	X		?	X
LRA	X		X	
Indel Calling	X	X	X	

At the time of analysis, the publication for Platypus was not available, so not all features were able to be determined.

ROC curves for all variants

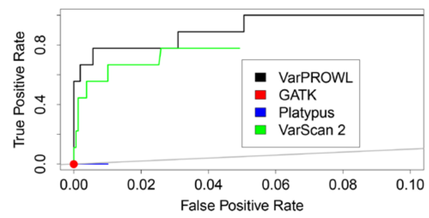
ROC curves for positions with variant frequency > 1% on QCCP that have been orthogonally validated on ASCP as TP or TN.

VarPROWL has the best performance, with a high true positive rate and a low false positive rate.



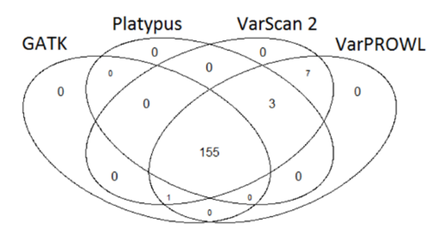
ROC curves among variants < 5%

GATK and Platypus never called variants below 5%. GATK also never called false positives.



Concordance between callers validated variants

GATK and Platypus never called variants below 5%. GATK also never called false positives.



Concordance between callers – QCCP

Concordance across the entire QCCP panel (not just orthogonally validated positions)

Program-specific variant calls were higher among callers better at detecting LFSNVs