Computational identification of within individual contamination for more sensitive somatic mutation profiling

Sangwoo Kim, Kyowon Jeong, Kunal Bhutani, Hayan Lee, and Vineet Bafna

Department of Computer Science and Engineering, UCSD

sak042@cs.ucsd.edu, vbafna@cs.ucsd.edu

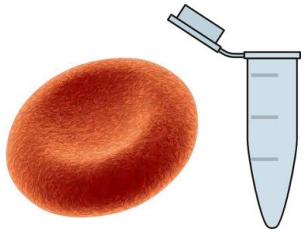
Contents

- Background
  - somatic mutation calling
  - **The Problem:** How to consider within individual contamination in somatic mutation calling
- Virmid:
  - **The solution:**
    - Estimate of sample heterogeneity (α)
    - Use α to call variants
  - Test & Validation
- Application:
  - Test on HME data set
  - New SNPs and their functions
- Summary and Conclusions

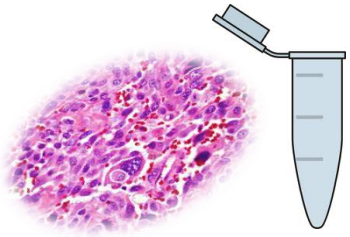
Variant calling and Within individual contamination

BACKGROUND

finding somatic mutations

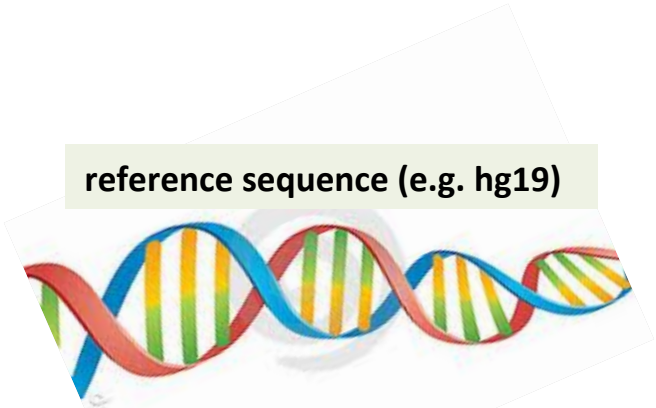


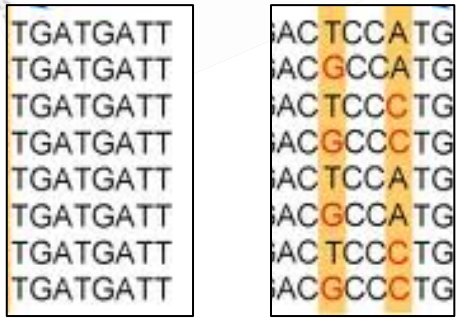
sample from
non-disease site



sample from
disease site

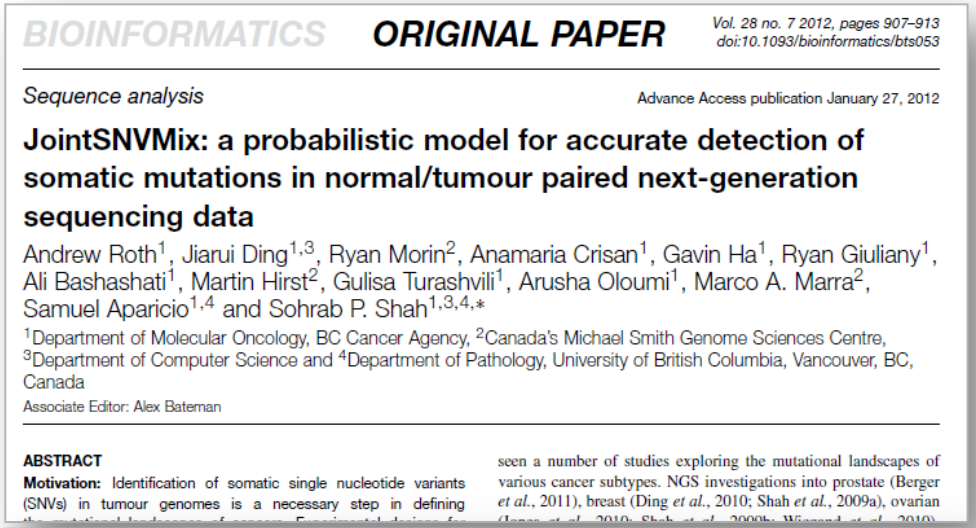
reference sequence (e.g. hg19)





- **UnifiedGenotyper**
- **VarScan**
- **SomaticSniper**
- **…**

Using a mixed model



BIOINFORMATICS **ORIGINAL PAPER** *Vol. 28 no. 7 2012, pages 907–913*
doi:10.1093/bioinformatics/bts053

Sequence analysis Advance Access publication January 27, 2012

JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data

Andrew Roth[1], Jiarui Ding[1,3], Ryan Morin[2], Anamaria Crisan[1], Gavin Ha[1], Ryan Giuliany[1], Ali Bashashati[1], Martin Hirst[2], Gulisa Turashvili[1], Arusha Oloumi[1], Marco A. Marra[2], Samuel Aparicio[1,4] and Sohrab P. Shah[1,3,4,*]

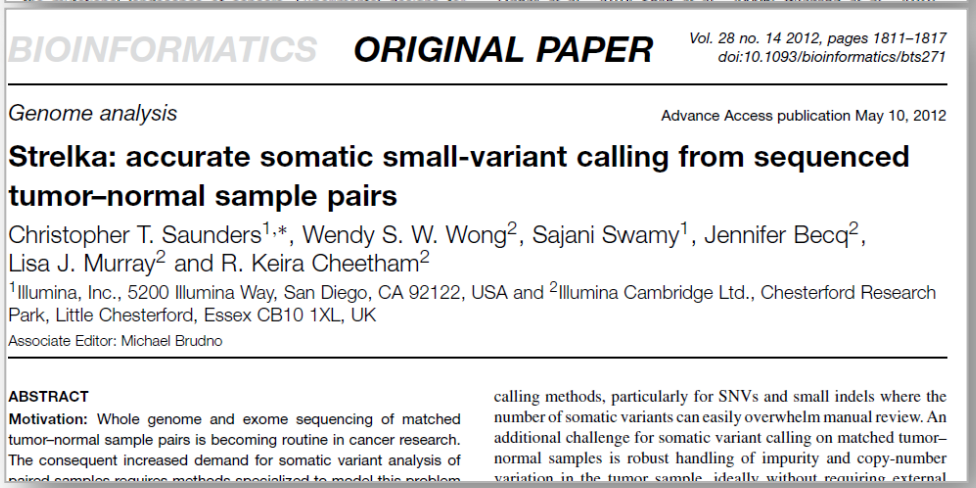
[1]Department of Molecular Oncology, BC Cancer Agency, [2]Canada's Michael Smith Genome Sciences Centre, [3]Department of Computer Science and [4]Department of Pathology, University of British Columbia, Vancouver, BC, Canada

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Identification of somatic single nucleotide variants (SNVs) in tumour genomes is a necessary step in defining

seen a number of studies exploring the mutational landscapes of various cancer subtypes. NGS investigations into prostate (Berger *et al.*, 2011), breast (Ding *et al.*, 2010; Shah *et al.*, 2009a), ovarian



BIOINFORMATICS **ORIGINAL PAPER** *Vol. 28 no. 14 2012, pages 1811–1817*
doi:10.1093/bioinformatics/bts271

Genome analysis Advance Access publication May 10, 2012

Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs

Christopher T. Saunders[1,*], Wendy S. W. Wong[2], Sajani Swamy[1], Jennifer Becq[2], Lisa J. Murray[2] and R. Keira Cheetham[2]

[1]Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA and [2]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL, UK

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Whole genome and exome sequencing of matched tumor–normal sample pairs is becoming routine in cancer research. The consequent increased demand for somatic variant analysis of

calling methods, particularly for SNVs and small indels where the number of somatic variants can easily overwhelm manual review. An additional challenge for somatic variant calling on matched tumor–normal samples is robust handling of impurity and copy-number variation in the tumor sample, ideally without requiring external

normal $p(AA), p(AB), p(BB)$

disease $p(AA), p(AB), p(BB)$

disease

$AA\&AB\&BB$

normal

$(p\downarrow 11\ \&p\downarrow 12\ \&p\downarrow 13\ @p\downarrow 21\ \&p\downarrow 22\ \&p\downarrow 23\ @p\downarrow 31\ \&p\downarrow 32\ \&p\downarrow 33\ )$

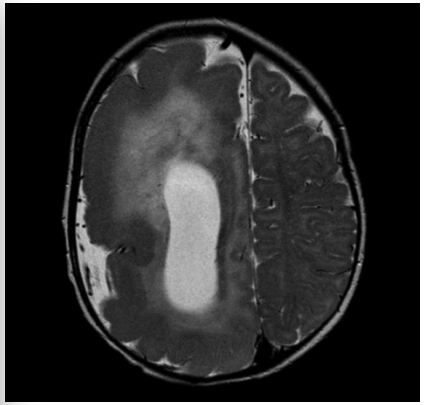
G: joint genotype probability matrix

LETTERS

nature genetics

De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly

Jeong Ho Lee[1,2,9], My Huynh[3,4], Jennifer L Silhavy[1,2], Sangwoo Kim[5], Tracy Dixon-Salazar[1,2], Andrew Heiberg[1,2], Eric Scott[1,2], Vineet Bafna[5], Kiley J Hill[1,2], Adrienne Collazo[1,2], Vincent Funari[6,7], Carsten Russ[8], Stacey B Gabriel[8], Gary W Mathern[3,4,10] & Joseph G Gleeson[1,2,10]



Focal cortical dysplasia

PNAS

Universal noninvasive detection of solid organ transplant rejection

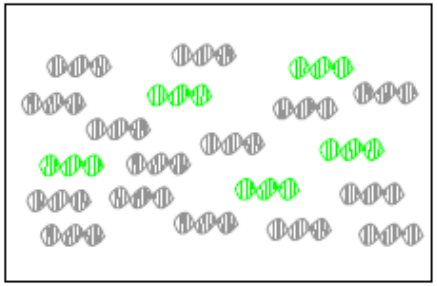
Thomas M. Snyder[a,b], Kiran K. Khush[c], Hannah A. Valantine[c,1], and Stephen R. Quake[a,b,1]

[a]The Howard Hughes Medical Institute and [b]Departments of Applied Physics and Bioengineering, Stanford University, Stanford, CA 94305; and [c]Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94305

Edited* by Leonard A. Herzenberg, Stanford University, Stanford, CA, and approved February 24, 2011 (received for review September 15, 2010)

It is challenging to monitor the health of transplanted organs, particularly with respect to rejection by the host immune system. Because transplanted organs have genomes that are distinct from the recipient's genome, we used high throughput shotgun sequencing to develop a universal noninvasive approach to monitoring organ health. We analyzed cell-free DNA circulating in the blood of heart

donor-specific chromosome Y has been detected in recipient urine and plasma (16–19). To date, most measurements of cell-free DNA in organ transplantation have been limited to the special case of women who receive male organs, which has prevented the widespread use of cell-free DNA as a diagnostic tool, because female recipients of male donor organs represent less than a quarter of all transplant procedures. HLA markers can be quantified to identify



Cell free DNA in recipients' blood

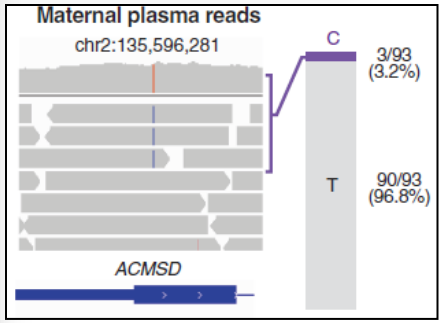
RESEARCH ARTICLE

GENOMICS

Noninvasive Whole-Genome Sequencing of a Human Fetus

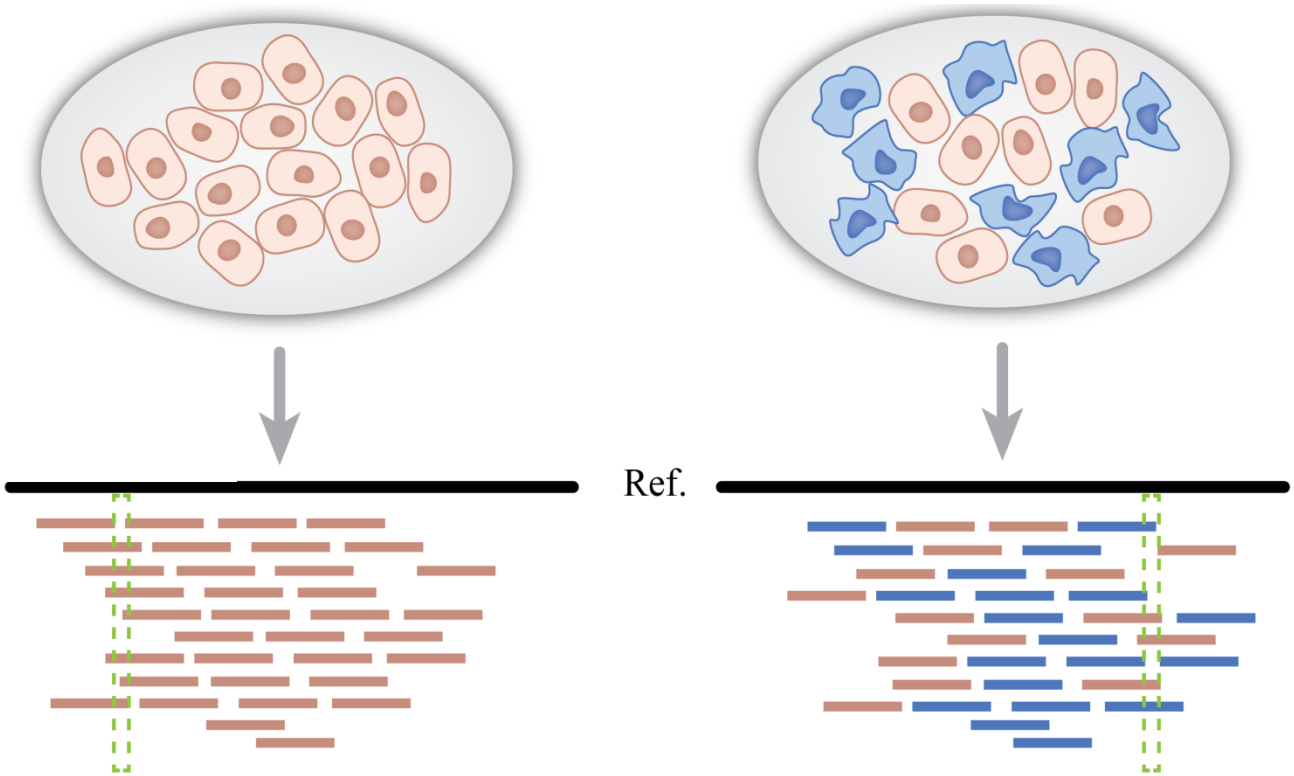
Jacob O. Kitzman,[1*] Matthew W. Snyder,[1] Mario Ventura,[1,2] Alexandra P. Lewis,[1] Ruolan Qiu,[1] LaVone E. Simmons,[3] Hilary S. Gammill,[3,4] Craig E. Rubens,[5,6] Donna A. Santillan,[7] Jeffrey C. Murray,[8] Holly K. Tabor,[5,9] Michael J. Bamshad,[1,5] Evan E. Eichler,[1,10] Jay Shendure[1*]

Analysis of cell-free fetal DNA in maternal plasma holds promise for the development of noninvasive prenatal genetic diagnostics. Previous studies have been restricted to detection of fetal trisomies, to specific paternally inherited mutations, or to genotyping common polymorphisms using material obtained invasively, for example, through chorionic villus sampling. Here, we combine genome sequencing of two parents, genome-wide maternal haplotyping, and deep sequencing of maternal plasma DNA to noninvasively determine the genome sequence of a human fetus at 18.5 weeks of gestation. Inheritance was predicted at 2.8 × 10^6 parental heterozygous sites

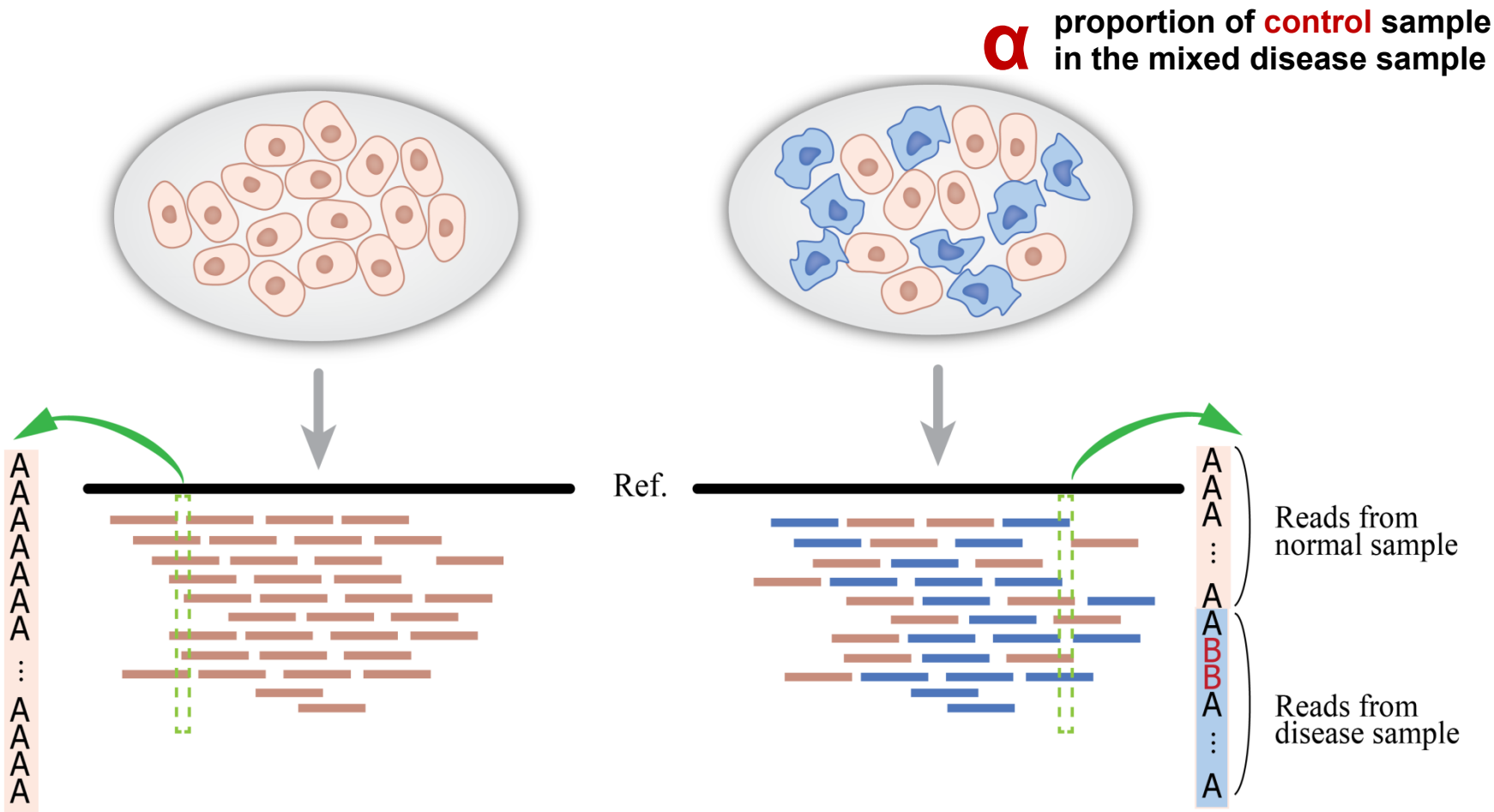


Fetus genome in maternal blood

Sample heterogeneity



Loss of somatic mutation calling

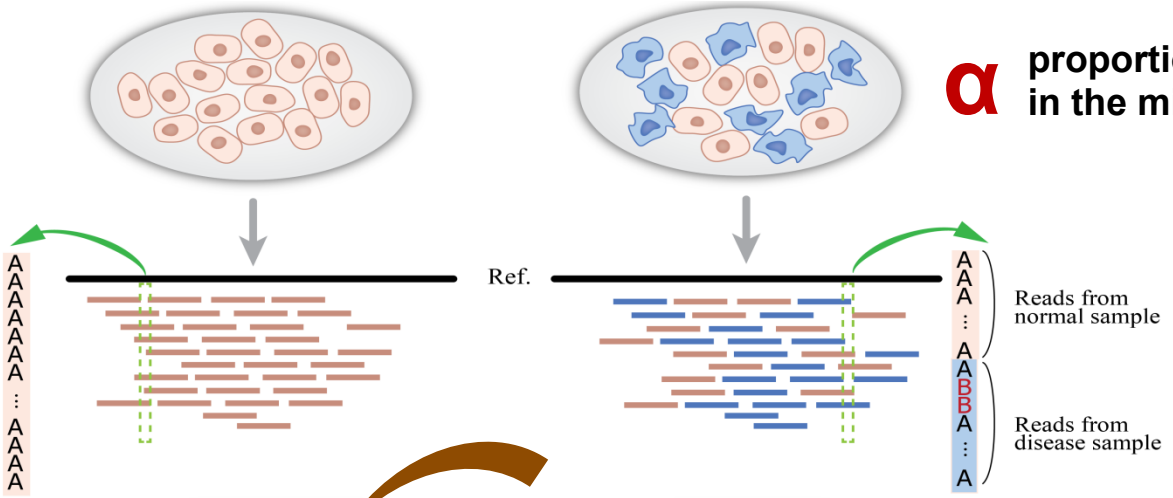


α=0 $g(AA,AB) \rightarrow ...ABBBABBAAAB... \rightarrow \sim 50\%\ BAF$

α=0.5 $g(AA,AB) \rightarrow ...AAAAABBAAAB... \rightarrow \sim 25\%\ BAF$

Virtual Microdissection for mixed disease sample

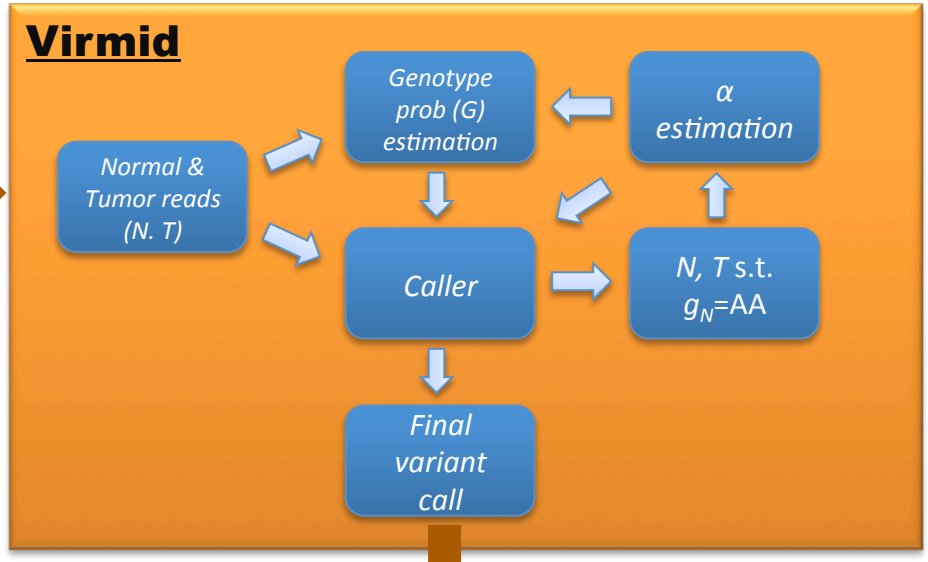
VIRMID



α **proportion of control sample in the mixed disease sample**

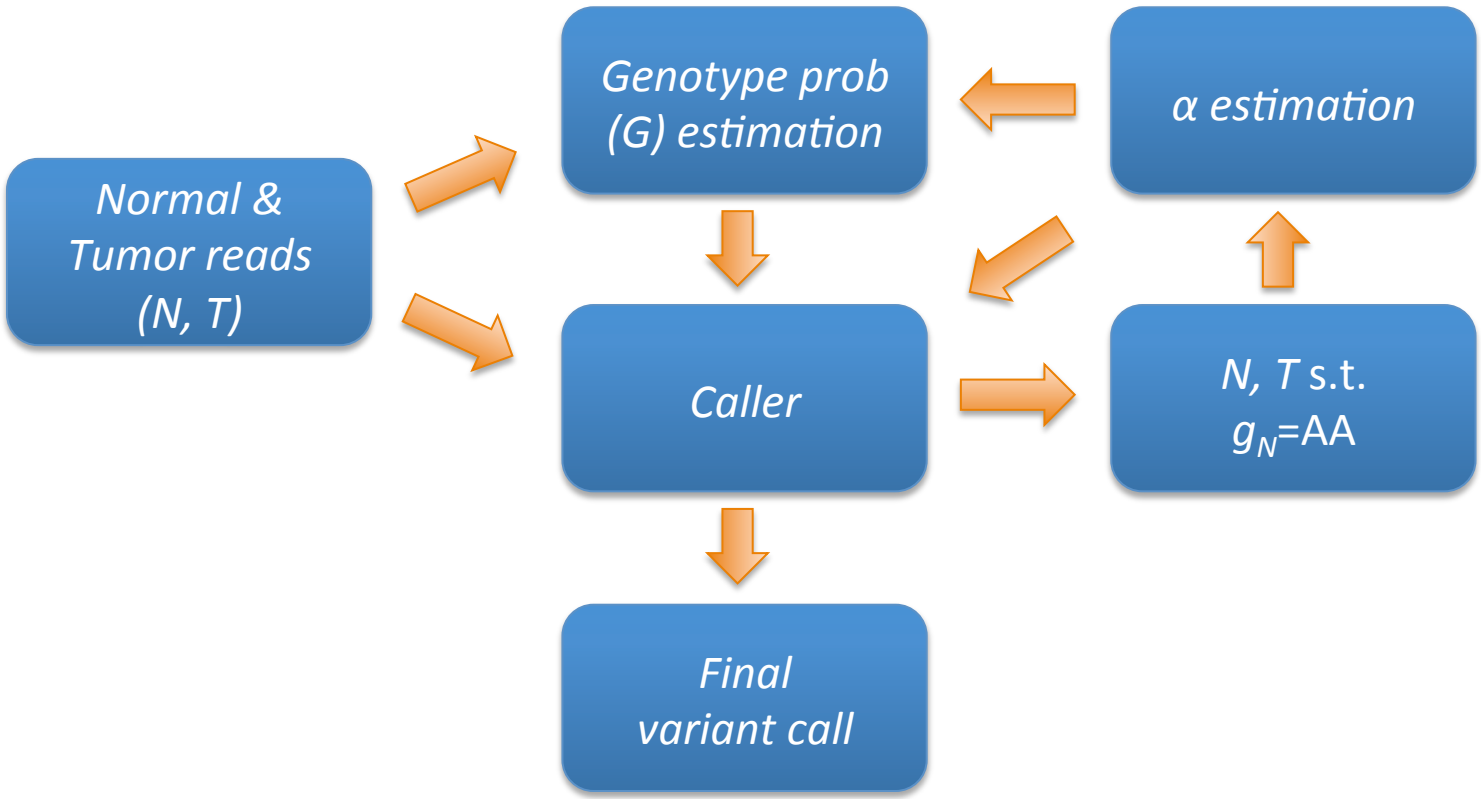
post-alignment process (GATK):
IndelRealigner, QualityRecalibration

1. **read alignment of disease & normal**
2. **reference sequence**



1. **estimated α**
2. **a list of somatic (germline) mutations**

Virmid Model



Likelihood function

- Likelihood function

$$L(\alpha, G \mid N, T) = P_{\alpha,G}(N,T)$$

$$= \prod_i \sum_{g_N, g_T} P_{\alpha,G}(g_N, g_T) \cdot \prod_j P_{\alpha,G}(N_j^i \mid g_N) \cdot \prod_k P_{\alpha,G}(T_k^i \mid g_N, g_T)$$

Joint genotype probability

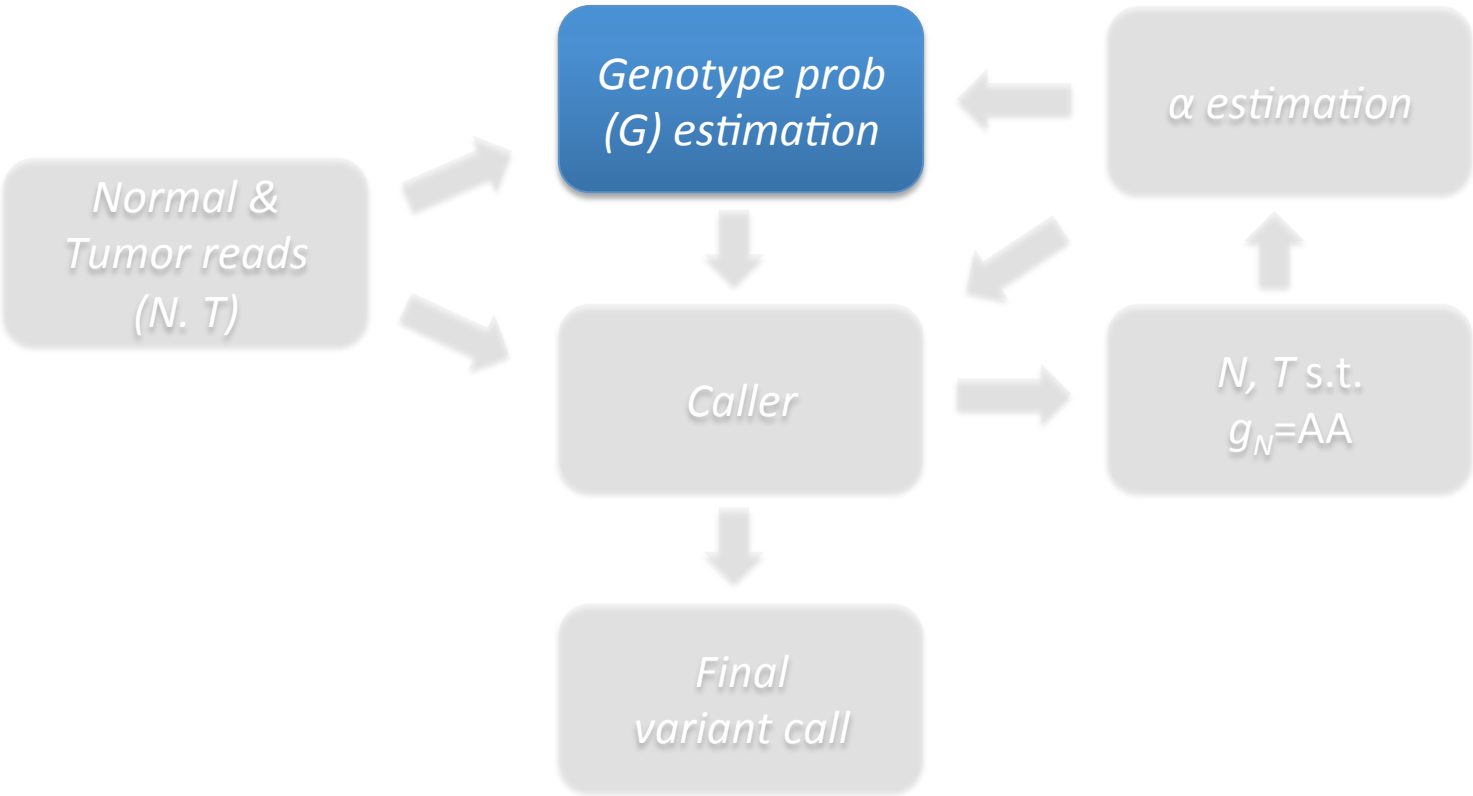
Probability that we observe the normal reads given the genotype of normal sample

Probability that we observe the tumor reads given the genotypes of normal/tumor sample

- Assumptions
  - Reads at different positions are independent
  - If genotypes are given, reads in a positions are independent

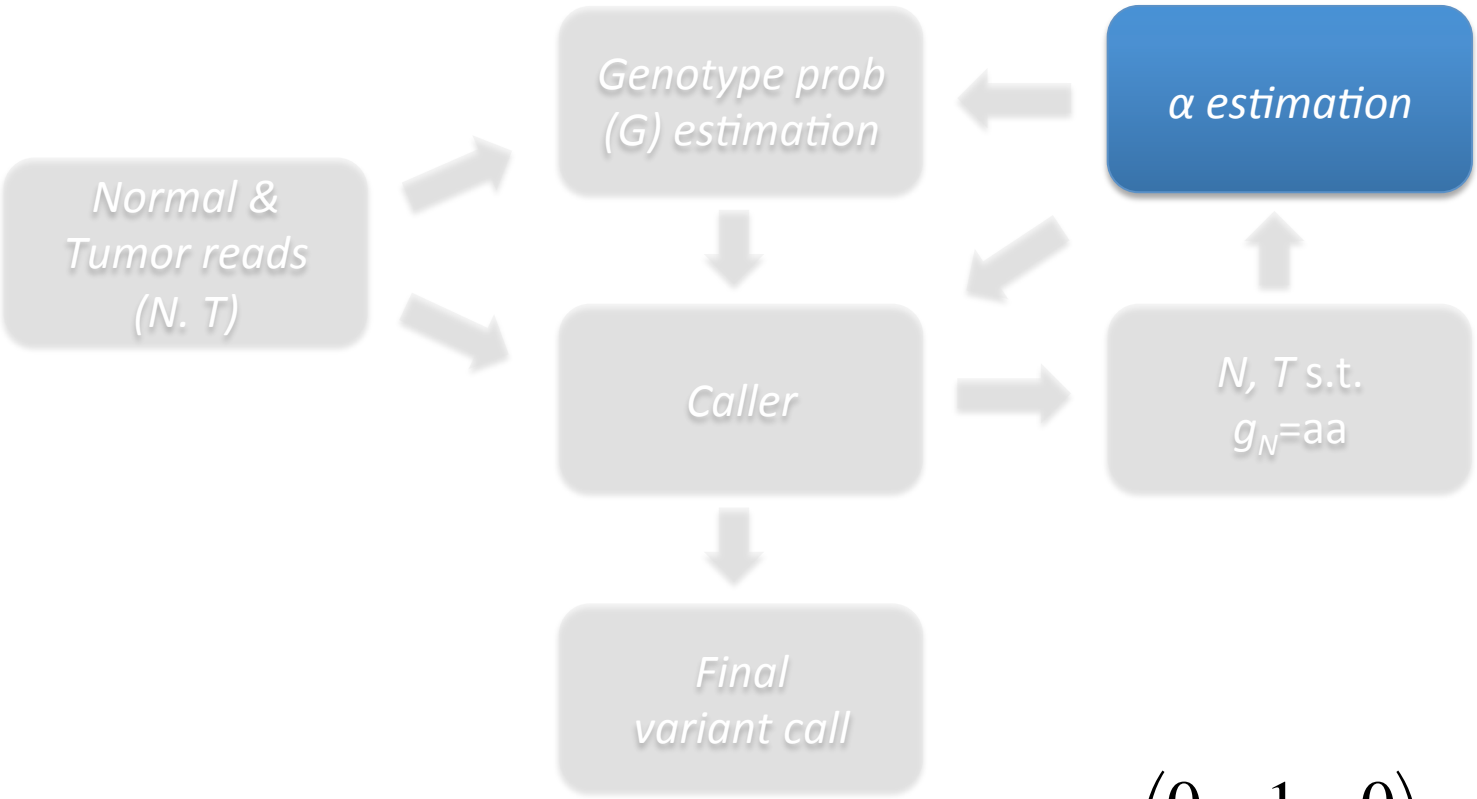
- Considers
  - Read error rate
  - Mapping error rate
  - Other biases

Estimation of genotype probabilities



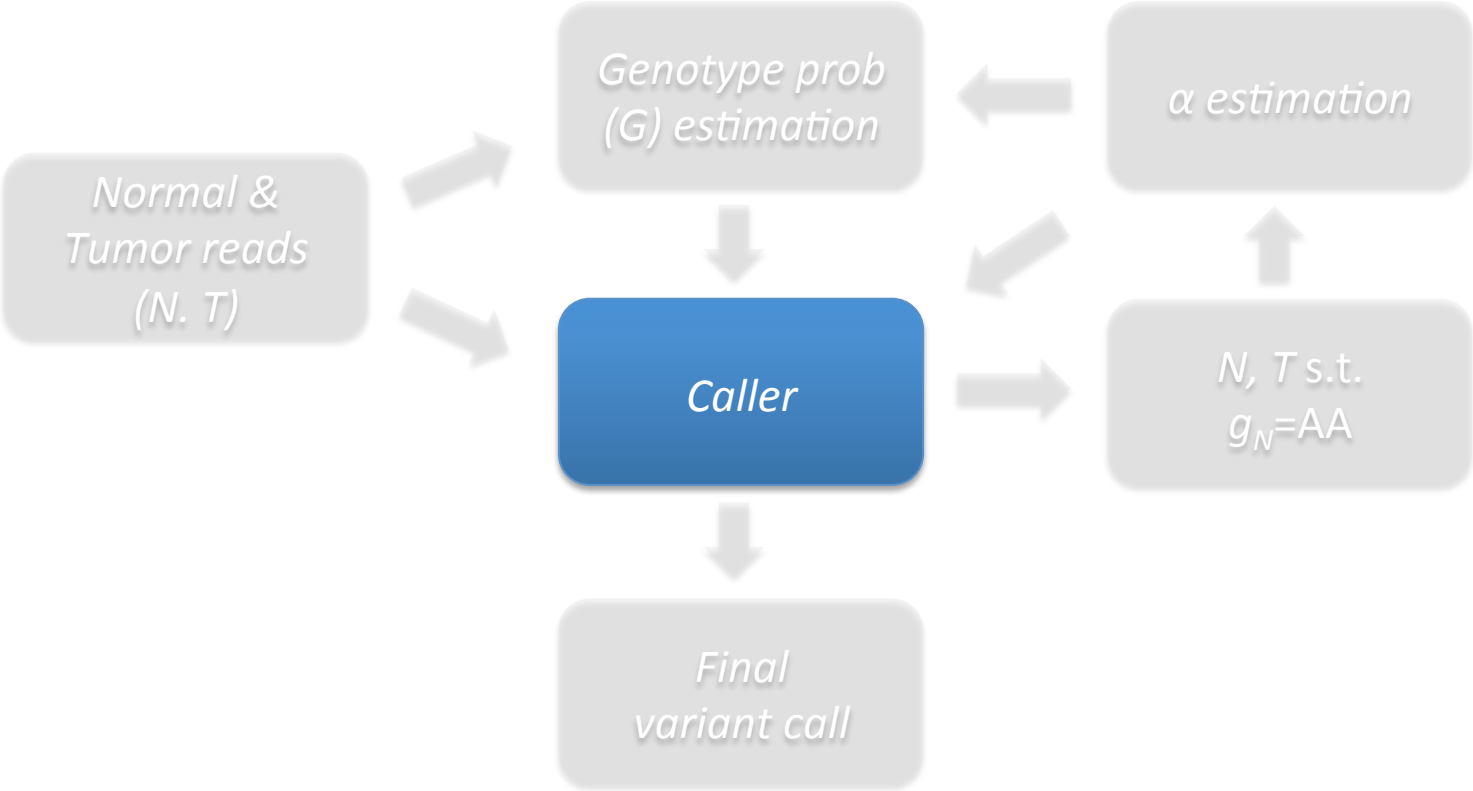
$$\hat{G} = \arg\max_{G} L(\hat{\alpha}, G \mid N, T)$$

Estimation of α



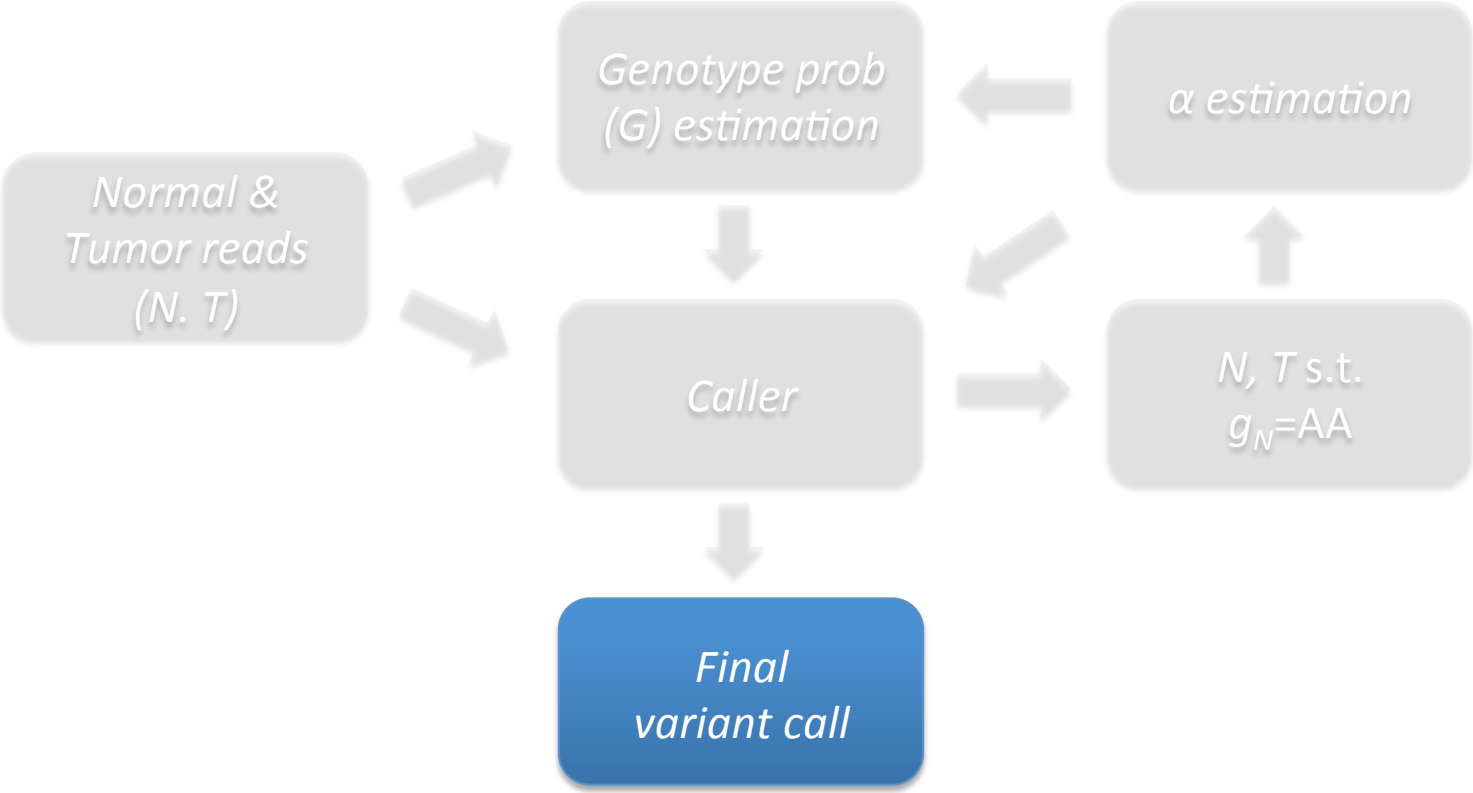
$$\hat{\alpha} = \arg\max_{\alpha} L(\alpha, \overline{G} \mid N, T) \quad \text{where} \quad \overline{G} \approx \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Variant Caller



- Use of MAP (Maximum a posteriori probability estimate)
- Use estimated *G* as a priori distribution
- Estimate allele frequencies

Final Variant Call



- Output final variants after α converged
- Call if the probability of being somatic mutation is higher than not

Validation

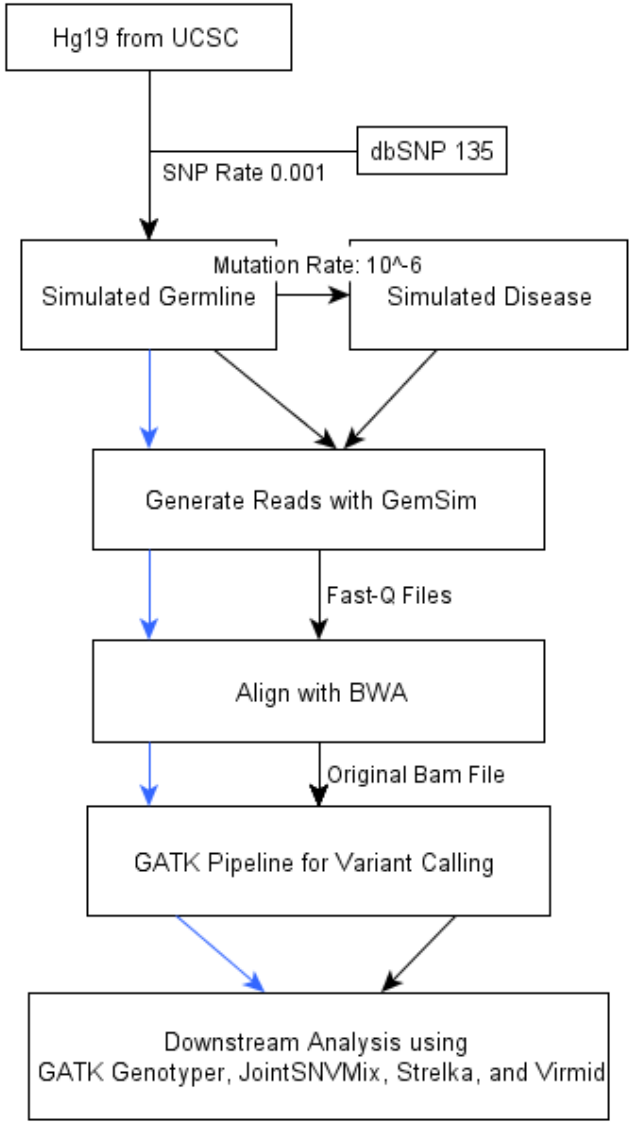
simulated mutation & mixture in hg19 Chr 1

tested α values:

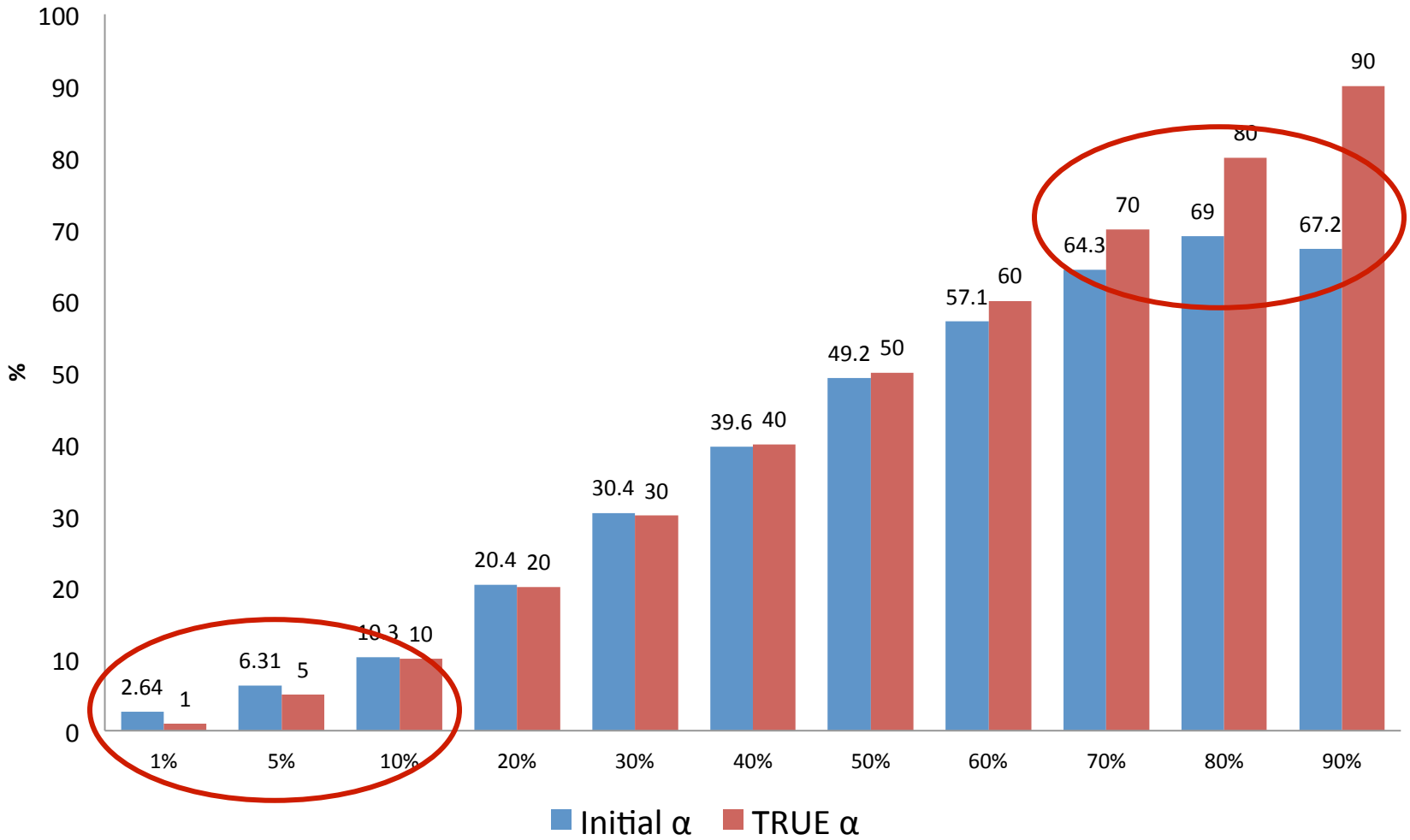
1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%

 # true somatic (germline) mutations:

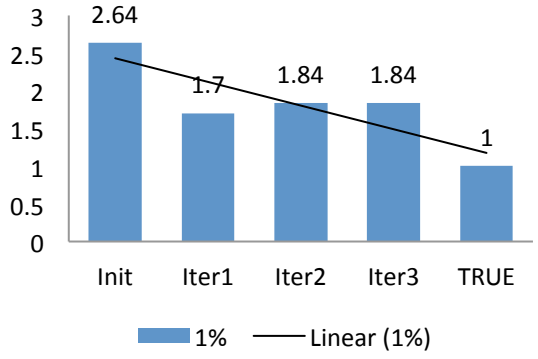
2226 (228,018)

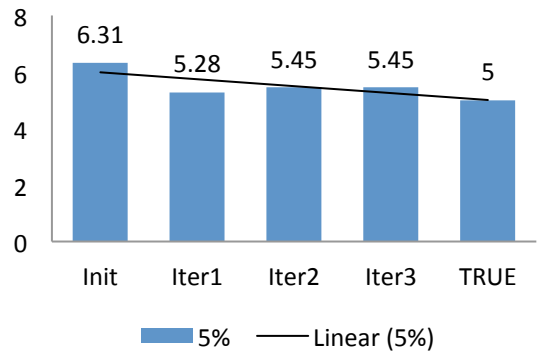


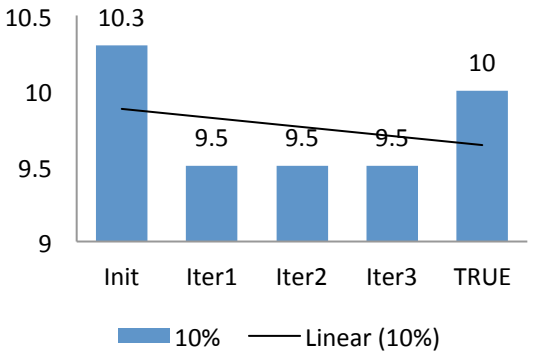
Estimation of α



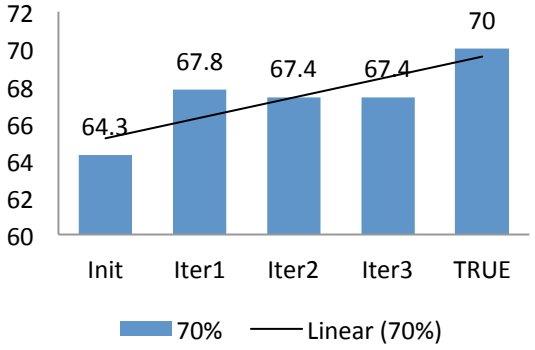
Reducing biases

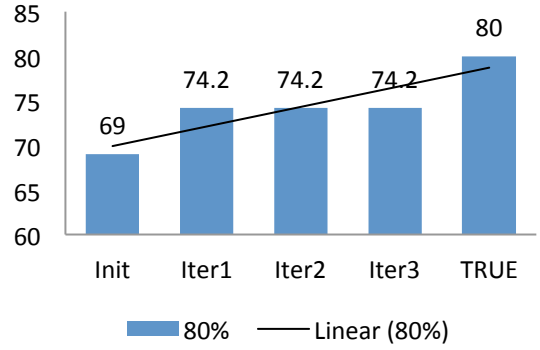


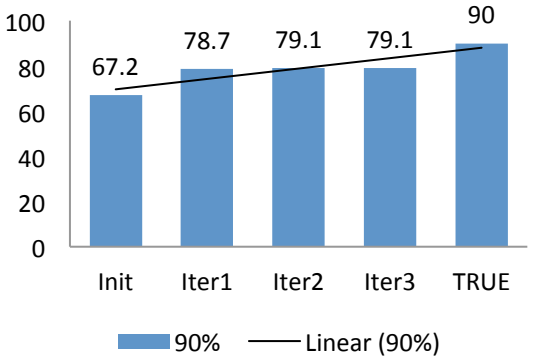




Changes in low α

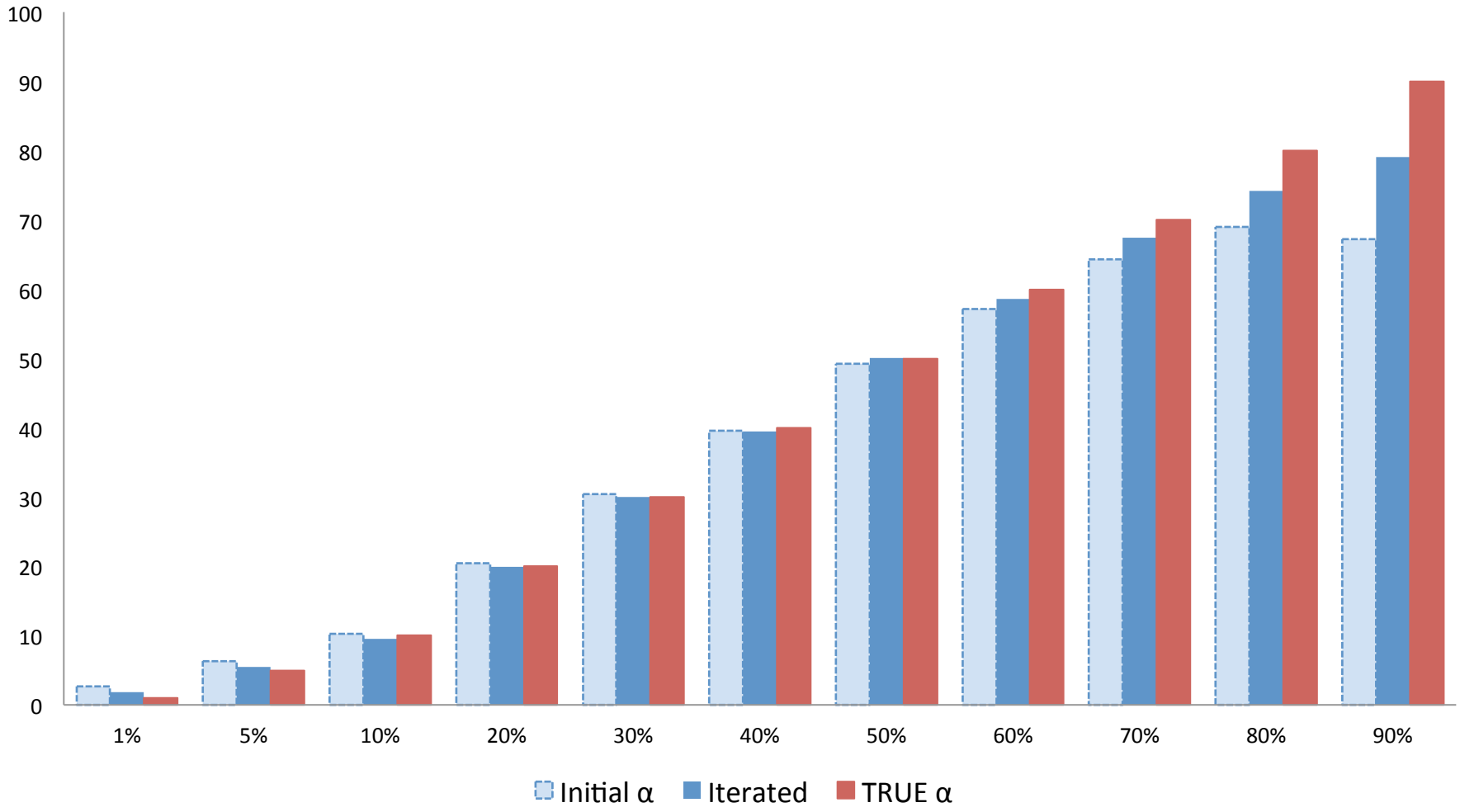






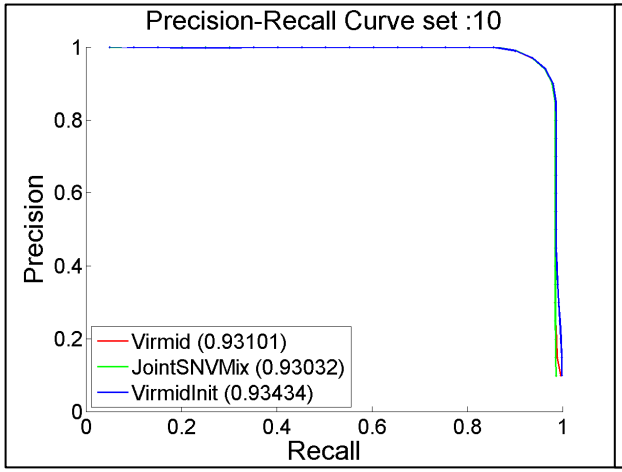
Changes in high α

α after iteration

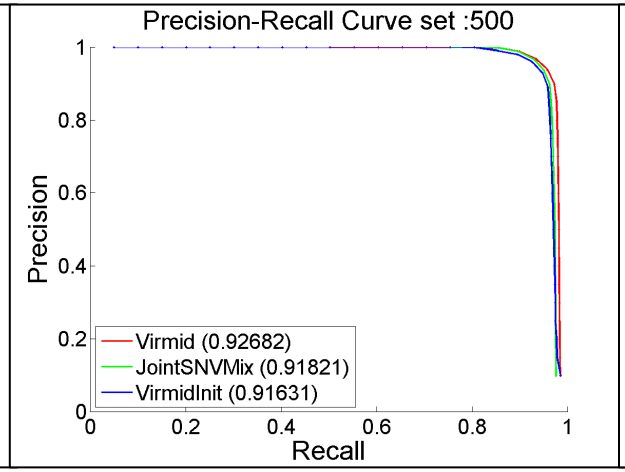


PR (Precision-Recall) Curves

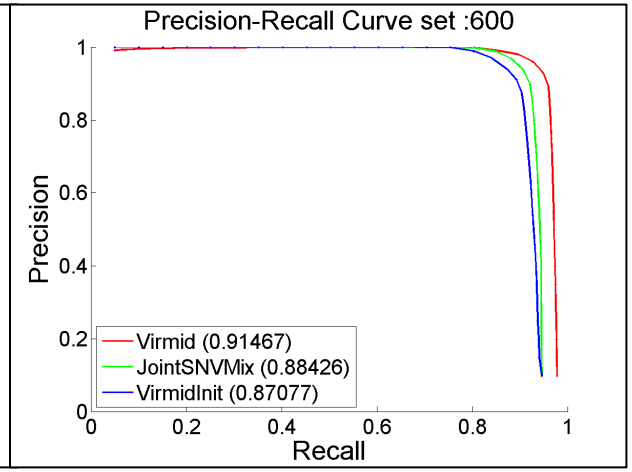
α=0.1%



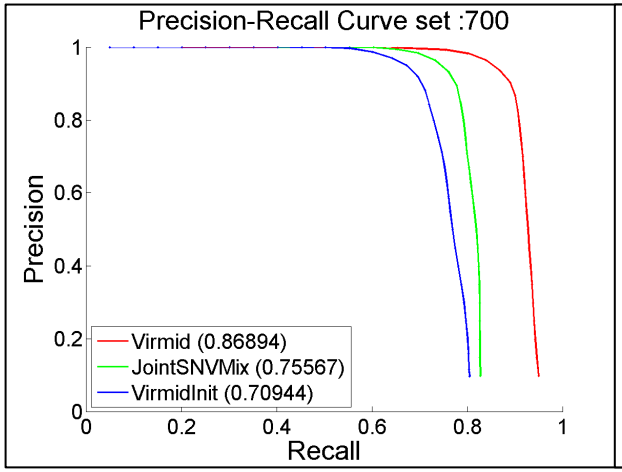
α=50%



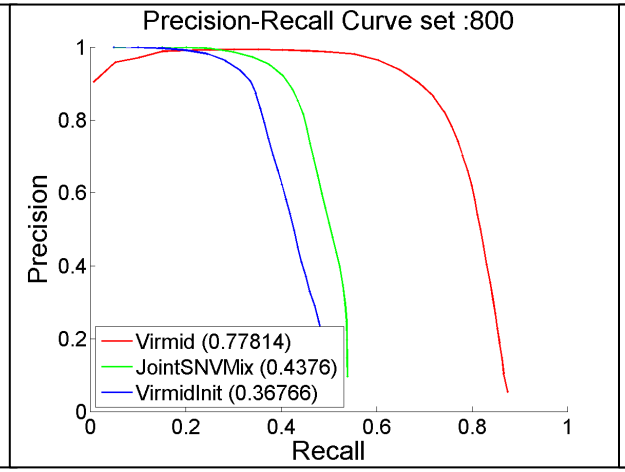
α=60%



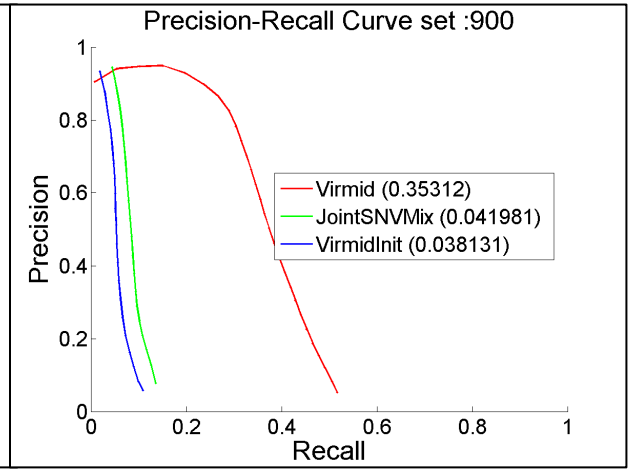
α=70%



α=80%

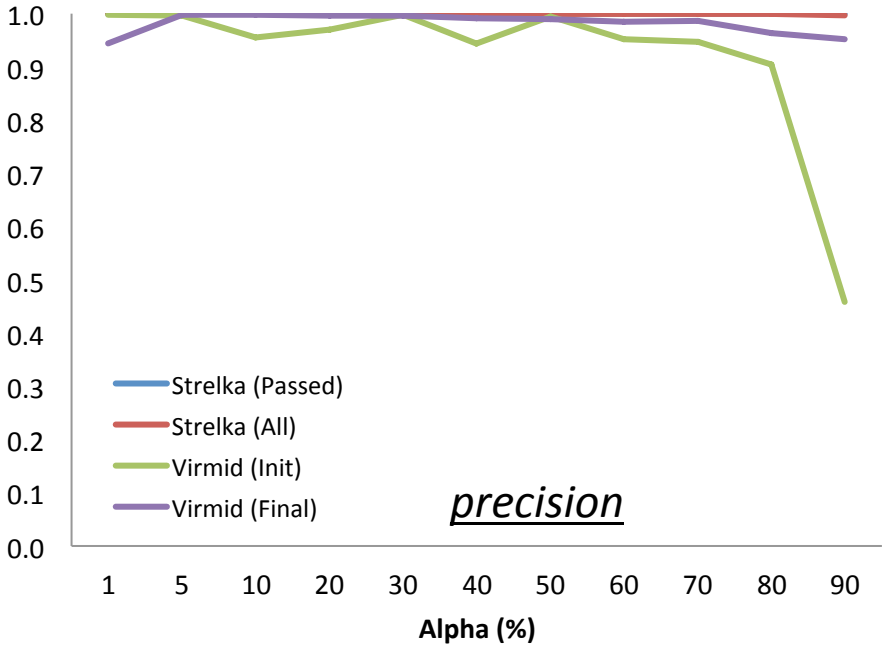


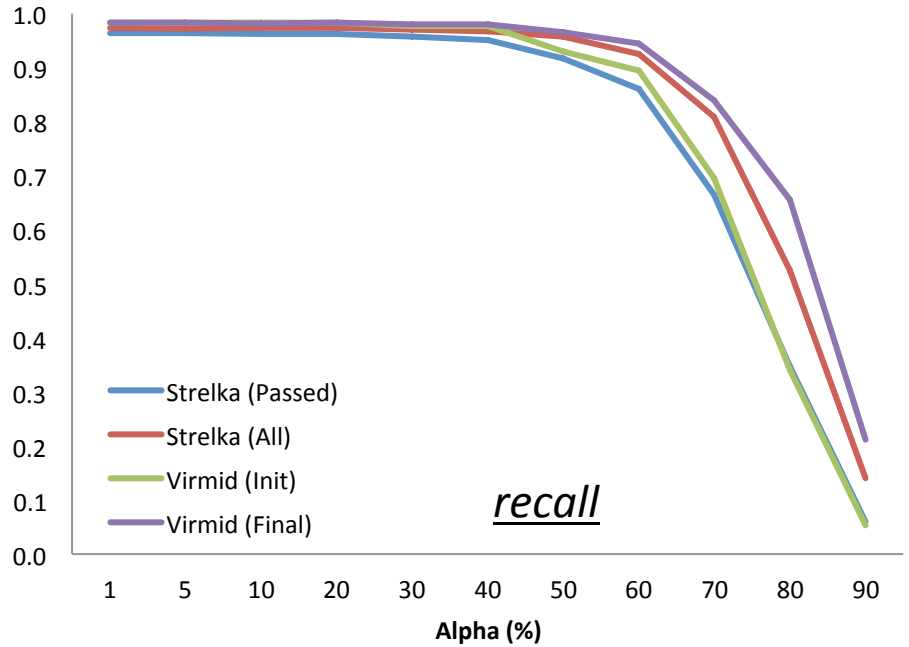
α=90%

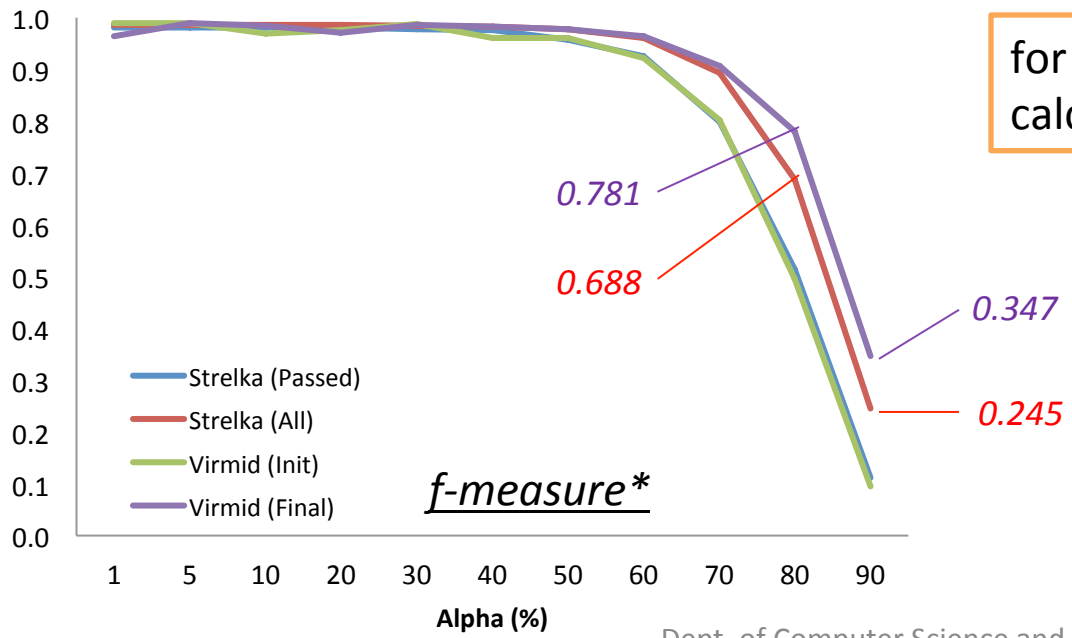


precision$=\#right/\#predicted=\#TP/\#TP+\#FP$

recall$=\#right/\#true=\#TP/\#TP+\#FN$







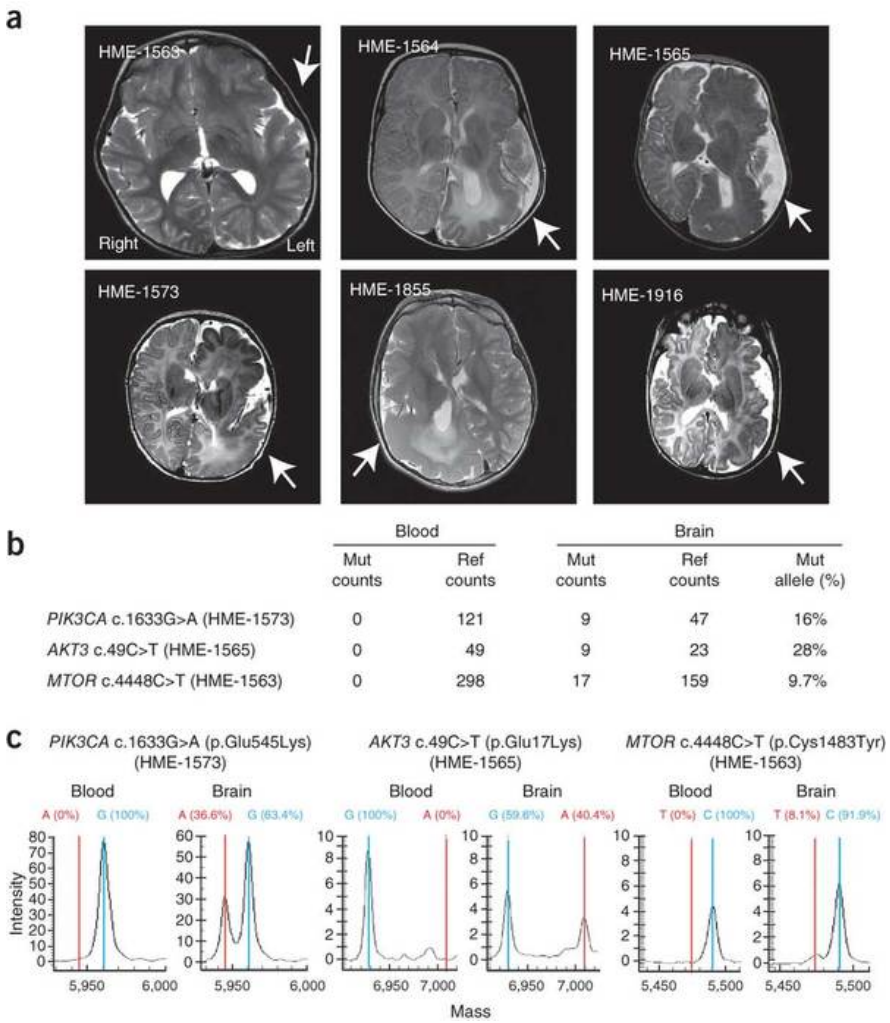
for each α, precision, recall, f-score* are calculated only from "final" outputs

$*f = 2 \times \text{precision} \times \text{recall} / \text{precision} + \text{recall}$

HME data

APPLICATION

HME-data



	Blood		Brain		
	Mut counts	Ref counts	Mut counts	Ref counts	Mut allele (%)
PIK3CA c.1633G>A (HME-1573)	0	121	9	47	16%
AKT3 c.49C>T (HME-1565)	0	49	9	23	28%
MTOR c.4448C>T (HME-1563)	0	298	17	159	9.7%

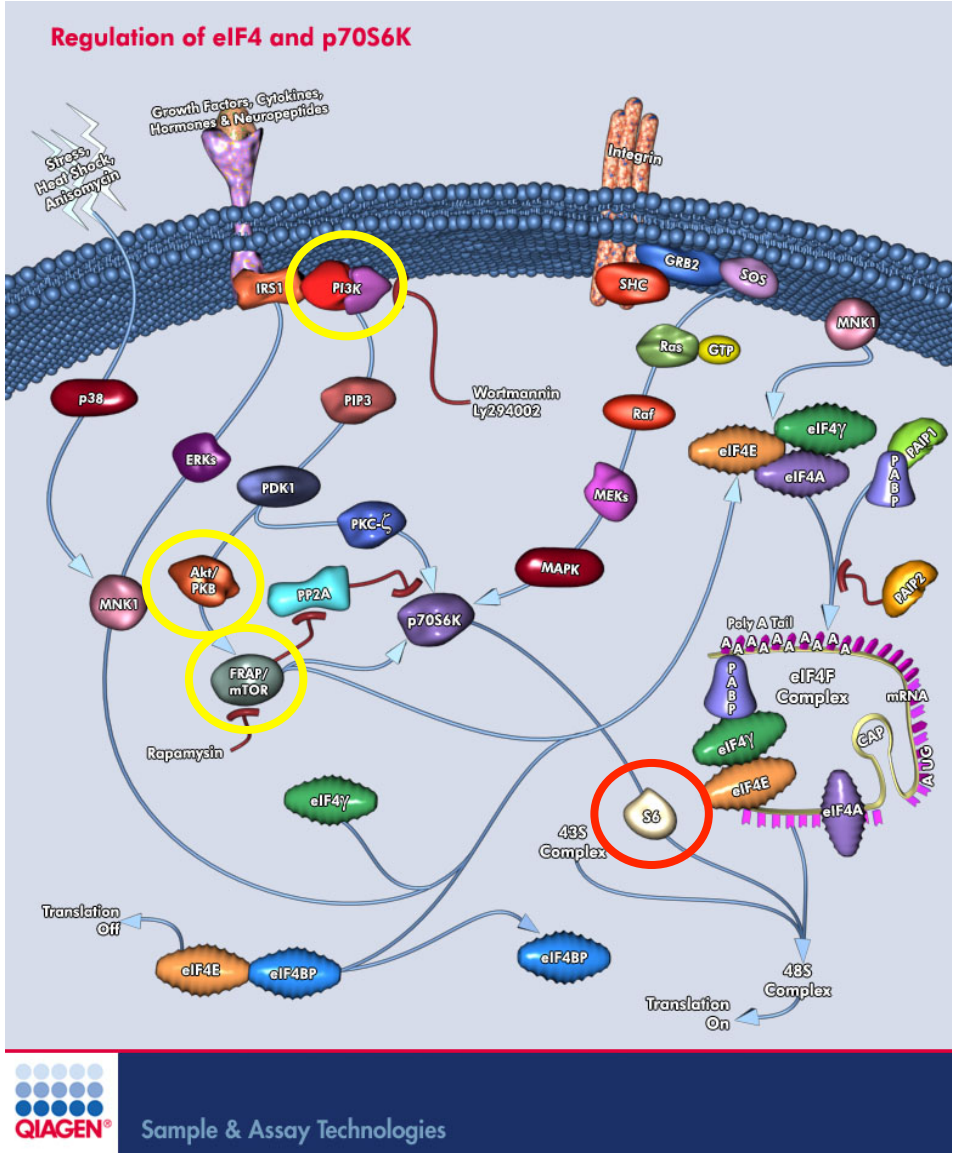
Sample	Previous Mut.	Mut. site
HME-1563	mTOR	Cys1483Tyr
HME-1565	AKT3	Glu17Lys
HME-1573	PIK3CA	Glu545Lys
HME-1574	-	-
HME-1620	-	-

SNPs called by JointSNVMix2

Variant calling with Virmid

Sample	Est. Alpha	Previous Calls	Virmid Calls	Misc.
HME-1563	**67.1%**	1547	1431	*mTOR in overlap*
HME-1565	**66.6%**	1328	1761	*AKT3 in overlap*
HME-1573	**62.8%**	1386	940	*PIK3CA in overlap*
HME-1574	**66.4%**	1440	3285	
HME-1620	**65.8%**	1335	4165	

More variants in a pathway



Chr	Pos	ref	sample	Func	AA Change	Gene
9	33796766	C	T	stop	GLN,stop	PRSS3
7	1.02E+08	G	A	missense	ARG,CYS	SPDYE6
3	25777564	G	T	missense	ASP,GLU	NGLY1
3	25777564	G	T	missense	ASP,GLU	NGLY1
3	25777564	G	T	missense	ASP,GLU	NGLY1
6	33410683	G	C	missense	ARG,PRO	SYNGAP1
6	30522401	A	C	missense	VAL,GLY	GNL1
18	29488748	A	G	missense	ILE,THR	TRAPPC8
10	1.14E+08	C	A	missense	CYS,PHE	GPAM
19	42795178	T	G	missense	VAL,GLY	CIC
1	16916476	G	C	missense	LEU,VAL	NBPF1
6	1.09E+08	C	G	missense	LEU,VAL	ARMC2
16	54967289	C	A	stop	SER,stop	IRX5
10	1.21E+08	T	C	splice-5	none	GRK5
5	71490955	C	A	missense	ASP,GLU	MAP1B

CONCLUSIONS

Conclusions

- Within individual contamination seriously affects somatic variant calling
- Virmid accurately infers the proportion of non-disease sample in a potentially mixed disease sample
- Virmid increases accuracy (precision and recall) by considering the within individual contamination
- By applying Virmid on disease samples with heterogeneity issues, we can identify more somatic variants to correlate with phenotypes

Acknowledgements

University of California, San Diego

Dept. of Computer Science and Engineering

Vineet Bafna, Ph.D.
Kunal Bhutani

Dept. of Electrical and Computer Engineering

Kyowon Jeong

Institute for Genomic Medicine, Rady Children's Hospital

Joseph Gleeson, M.D, Ph.D.
Jeong Ho Lee, M.D, Ph.D. (KAIST)

Dept. of Bioengineering

Hojung Nam, Ph.D.

Stony Brook University

Dept. of Computer Science

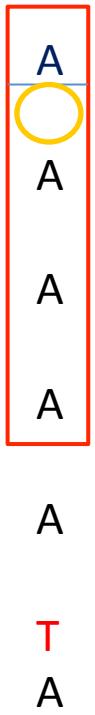
Hayan Lee

Funding:
P01 HD070494-01 and RO1 HG004962

Thank you

SUPPLEMENTARY SLIDES

Estimating calls



sequencing error = e

- True genotype = "AA" *and* no sequencing error
  - $P1-eg=AA\ P(g=AA)$
- True genotype = "AT"
  - Read was generated from 'A' allele *and* no sequencing error
  - $1/2 * P1-eg=AT\ P(g=AT)$
  - Read was generated from 'T' allele *and* sequencing error *and* 'A' was generated by chance
  - $1/2 * 1/4 * Peg=AT\ P(g=AT)$
- True genotype = "TT" *and* sequencing error
  - $Peg=TT\ P(g=TT)$

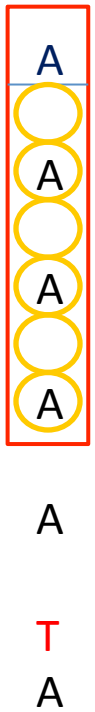
Estimating calls (cont'd)

A
A
A
A
A
T
A

sequencing error = e

- True genotype = "AA" *and* sequencing error
  - $Peg{=}AA\ P(g{=}AA)$
- True genotype = "AT"
  - Read was generated from 'A' allele *and* sequencing error *and* 'T' was generated by chance
  - $1/2 * 1/4 * Peg{=}AT\ P(g{=}AT)$
  - Read was generated from 'T' allele *and* no sequencing error
  - $1/2 * P1{-}eg{=}AT\ P(g{=}AT)$
- True genotype = "TT" *and* no sequencing error
  - $Peg{=}TT\ P(g{=}TT)$

Estimating calls (Cont'd)



$$P(AA|D) = PDAA\ P(AA) / PDAA\ P(AA) + PDAB\ P(AB) + PDBB\ P(BB)$$

$$= PDAA\ P(AA) / \sum gi \in G \uparrow G = \{AA, AT, TT\} \ PDgi\ P(gi)$$

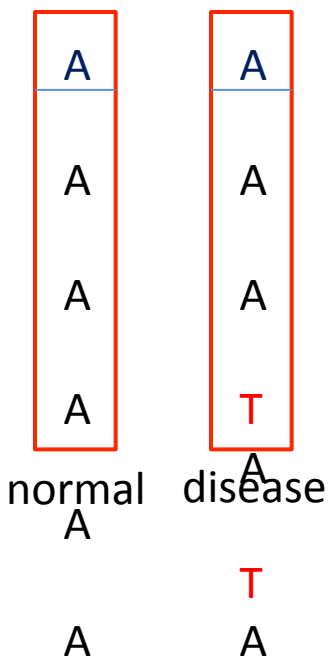
$$P(AB|D) = \ldots$$

$$P(BB|D) = \ldots$$

	G=AA	G=AB	G=BB
Probability	56%	40%	4%

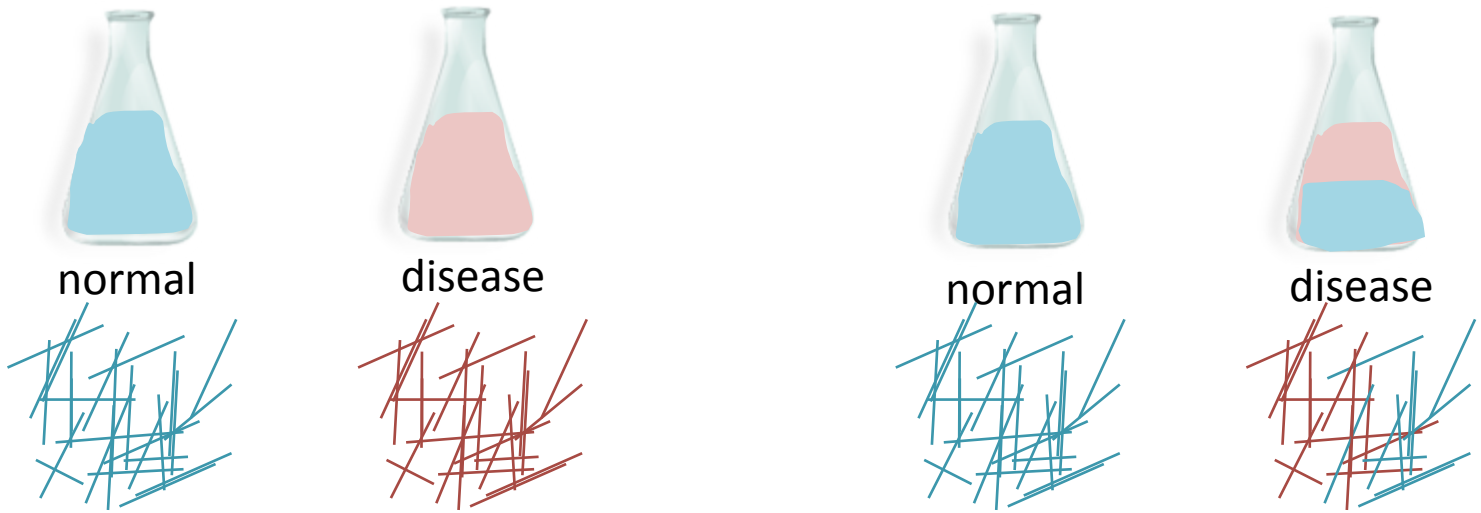
Calling somatic mutations

- Call variants for disease sample
  - And filter out if there's a same variant in normal sample (GATK-UnifiedGenotyper)
- Calculate a joint probability for totally 9 (3 x 3) genotypes (jointSNVMix, Strelka)



	G_T=AA	G_T=AB	G_T=BB
G_N=AA	10%	80%	2%
G_N=AB	0.5%	5%	1%
G_N=BB	0.1%	0.4%	1%

Within individual contamination

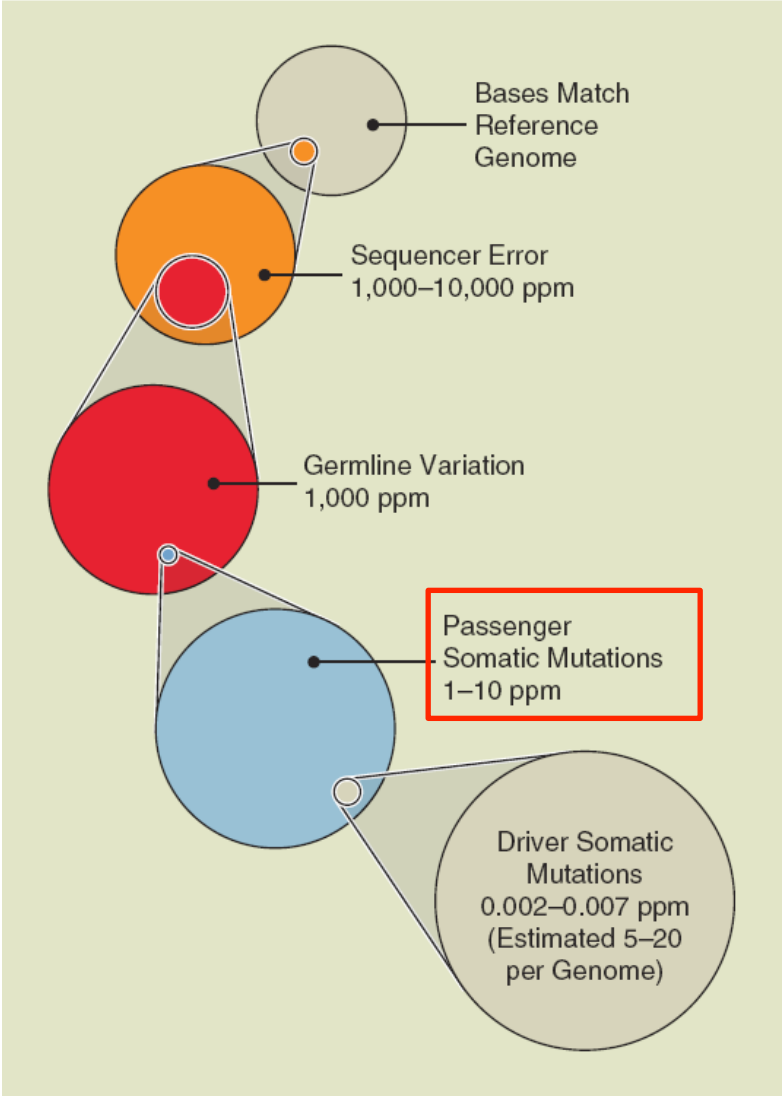


genotype = AA	AA	AB	BB
A	A	A	A
A	A	A	B
A	A	A	B
A	A	A	B
A	A	A	B
A	A	B	B
A	A	B	B
A	A	B	B
A	A	B	B

genotype = AA	AA	AB	BB
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	B
A	A	A	B
A	A	B	B
A	A	B	B
A	A	B	B

read from normal

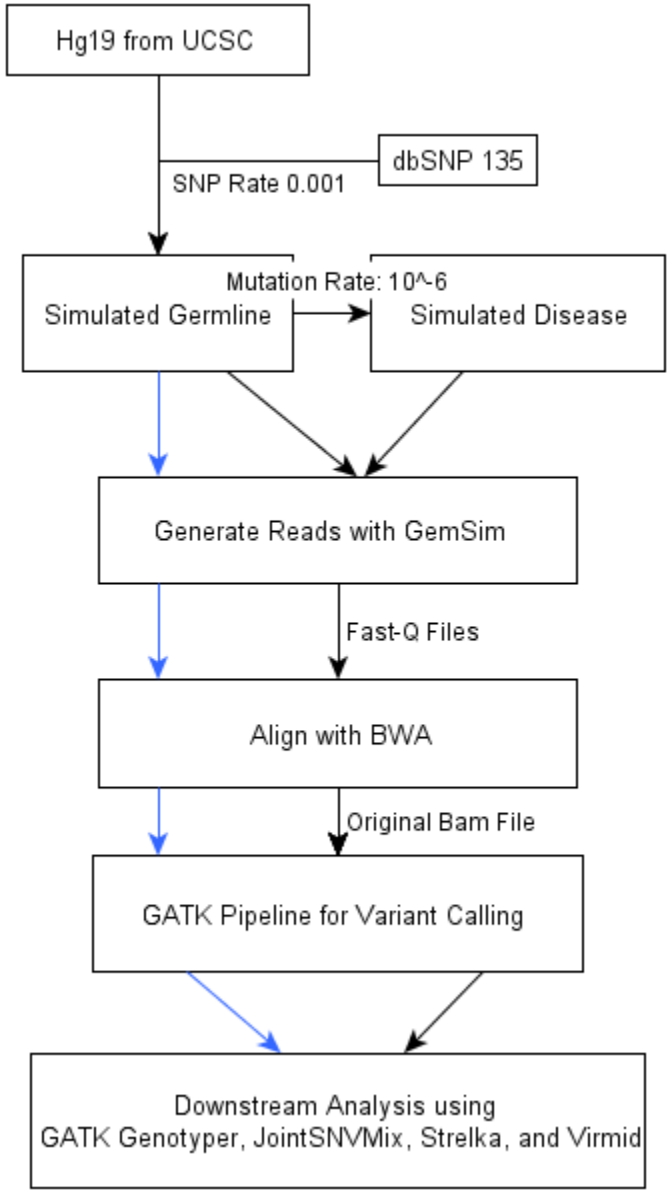
Somatic mutations are rare and difficult to find



Signal Processing Magazine, IEEE **Developing Algorithms to Discover Novel Cancer Genes: A look at the challenges and approaches**

Competitor Tools

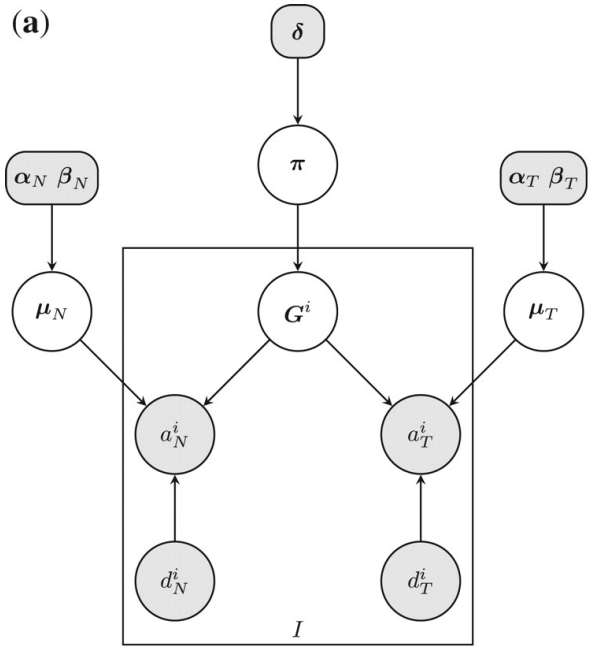
- Other NextGen Tools
  - VarScan
  - Somatic Sniper
- SNP-Array Tools
  - Absolute
  - OncoSNP



Workflow

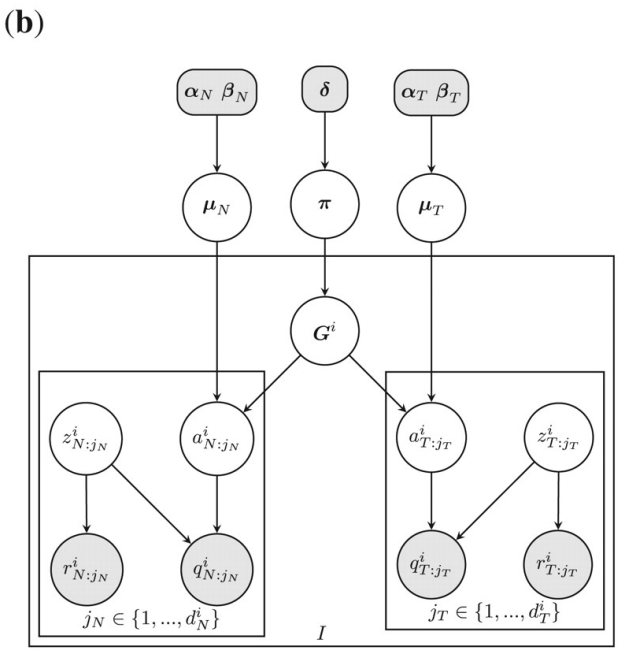
- GATK Unified Genotyper:
- JointSNVMix:
  - Default Settings with the joint_snv_mix_two option
- Strelka
  - Default settings with the bwa option

JointSNVMix1



$$\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\delta}) \\
\boldsymbol{G}^i|\boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{G}^i|\boldsymbol{\pi}) \\
a_N^i|G^i_{(g_N,g_T)}=1,\boldsymbol{\mu}_N,d_N^i &\sim \text{Binomial}(a_N^i|d_N^i,\mu_{N:g_N}) \\
\mu_{x:g}|\alpha_{x:g},\beta_{x:g} &\sim \text{Beta}(\mu_{x:g}|\alpha_{x:g},\beta_{x:g})
\end{aligned}$$

JointSNVMix2



$$\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\delta}) \\
\boldsymbol{G}^i|\boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{G}^i|\boldsymbol{\pi}) \\
a^i_{N:j_N}|G^i_{(g_N,g_T)}=1,\boldsymbol{\mu}_N &\sim \text{Bernoulli}(a^i_{N:j_N}|\mu_{N:g_N}) \\
z^i_{N:j_N} &\sim \text{Bernoulli}(z^i_{N:j_N}|0.5) \\
q^i_{N:j_N}|a^i_{N:j_N},z^i_{N:j_N} &\sim f(q^i_{N:j_N}|a^i_{N:j_N},z^i_{N:j_N}) \\
r^i_{N:j_N}|z^i_{N:j_N} &\sim g(r^i_{N:j_N}|z^i_{N:j_N}) \\
\mu_{x:g}|\alpha_{x:g},\beta_{x:g} &\sim \text{Beta}(\mu_{x:g}|\alpha_{x:g},\beta_{x:g})
\end{aligned}$$

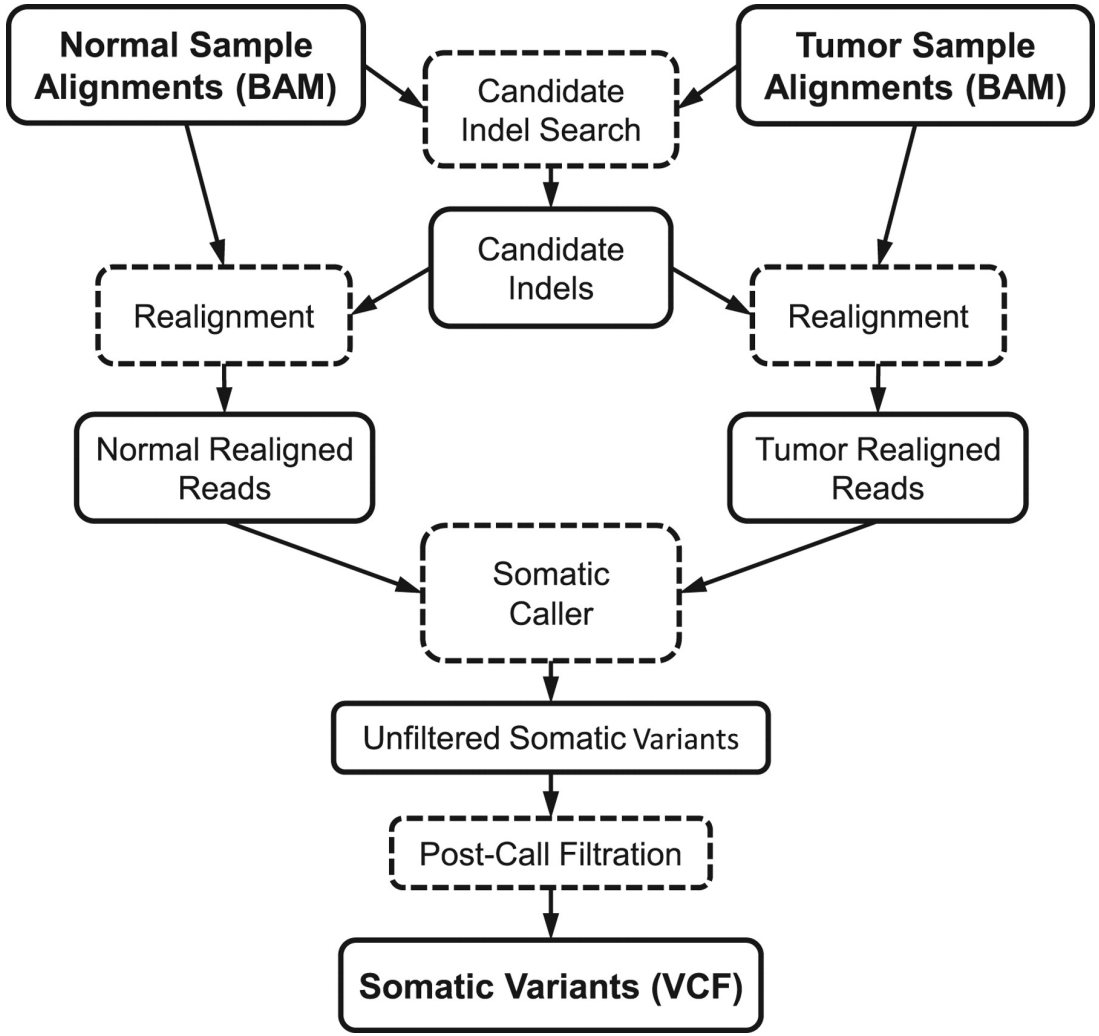
Roth A et al. Bioinformatics 2012;28:907-913

Strelka

- Developed at Illumina
- Maximizes posterior probability of the joint tumor and normal allele frequencies
  - Other tools only look at genotypes in terms of matching reference or not.
- Is able to call somatic variants as well as detect insertions and deletions
- Does not implicitly calculate contamination or tumor impurity
- High filtering: less false positives

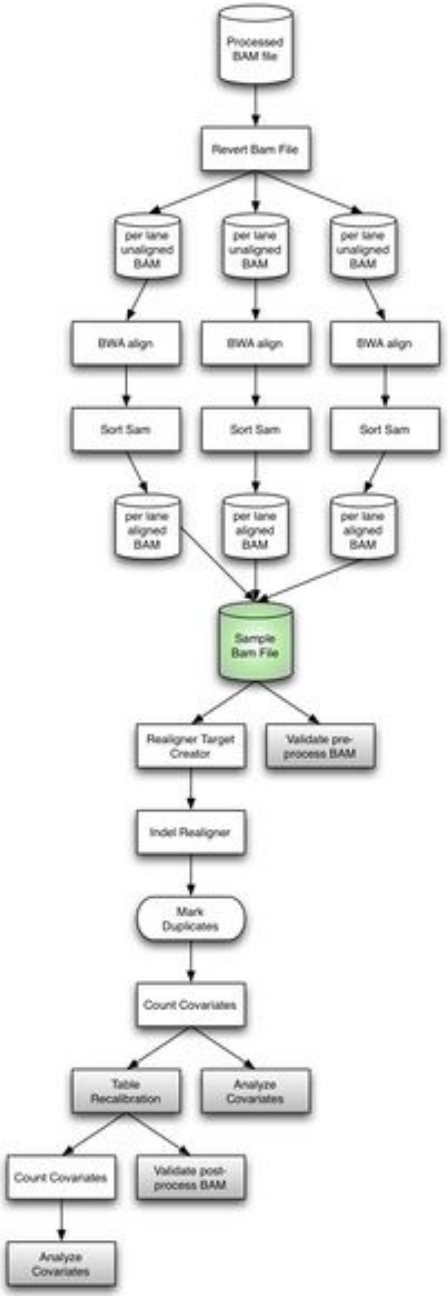
Saunders C T et al. Bioinformatics 2012;28:1811-1817

Strelka



Saunders C T et al. Bioinformatics 2012;28:1811-1817

GATK Best Practices

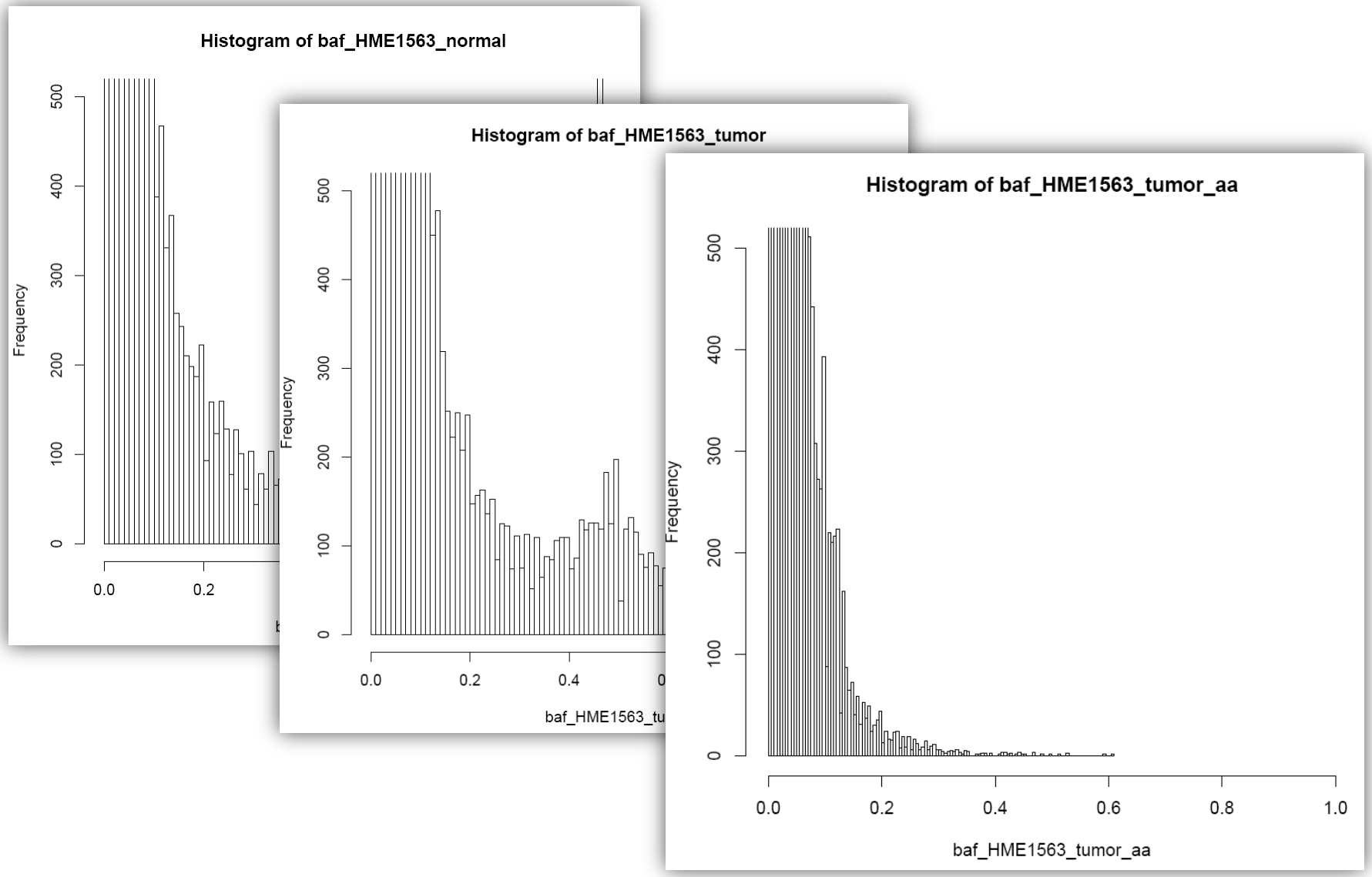


http://www.broadinstitute.org/gsa/wiki/index.php/Data_Processing_Pipeline

AUC (Area under curve)

Alpha (%)	JointSNVMix	Virmid(*init*)	Virmid	AUC diff.
1	0.930	**0.935**	**0.935**	
5	0.932	**0.936**	0.935	
10	0.929	0.934	**0.935**	$< 10^{-3}$
20	0.930	**0.933**	**0.933**	
30	0.926	0.930	**0.931**	
40	0.926	**0.931**	0.930	
50	0.918	0.915	**0.927**	**0.008**
60	0.884	0.870	**0.910**	**0.026**
70	0.756	0.702	**0.854**	**0.098**
80	0.438	0.356	**0.741**	**0.304**
90	N/A	0.040	**0.302**	N/A

WIC in HME data



Bias 1 - Loss of Reads

$$x_a = p(\text{a read that passes } g_1 \text{ being unmapped})$$
$$= p(r_1 \text{ has } d+1 \text{ or more variants in the remaining sites})$$

$$x_b = p(\text{a read that passes } g_2 \text{ being unmapped})$$
$$= p(r_2 \text{ has } d+1 \text{ or more variants in the remaining sites})$$

$$x_a = 1 - \sum_{i=0}^{d} \binom{l-1}{i} p^i (1-p)^{l-1-i}$$

$$x_b = 1 - \sum_{i=0}^{d-1} \binom{l-1}{i} p^i (1-p)^{l-1-i}$$

,where d=maximum edit distance, l=read length, and p=frequency of variation

Bias 2 - Loss of variants