Exome SNP validations

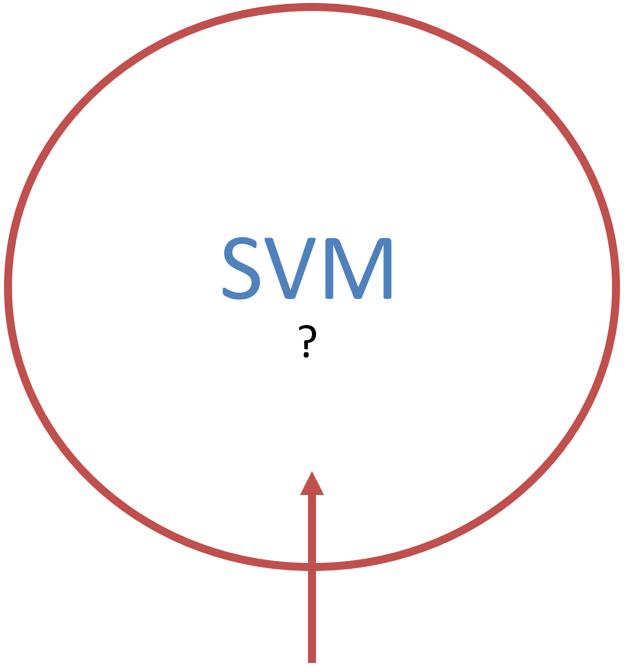
Jin Yu, Danny Challis, Aniko Sabo, Donna Muzny, Fuli Yu

Jan 2012

Overview

- Validation designs
  - Exome SNP consensus
    - overall: 200 sites stratified by AC on chr20
      - At most 5 samples picked from Oct 2011 official release integrated VCF
    - exclusive: 100 sites exclusive to VQSRv2b and dbSNP on chr20
      - At most 2 samples picked from Oct 2011 official release integrated VCF
  - Centers’ specific calls:
    - Centers’ unique sites not included in consensus, VQSRv2b and dbSNP, stratified by Illumina and SOLiD
      - At most 2 samples picked from individual call sets
- Results
  - Exome consensus sites quality
  - integrated genotypes quality
  - Center specific unselected sites quality

Validation design (1): SVM overall quality

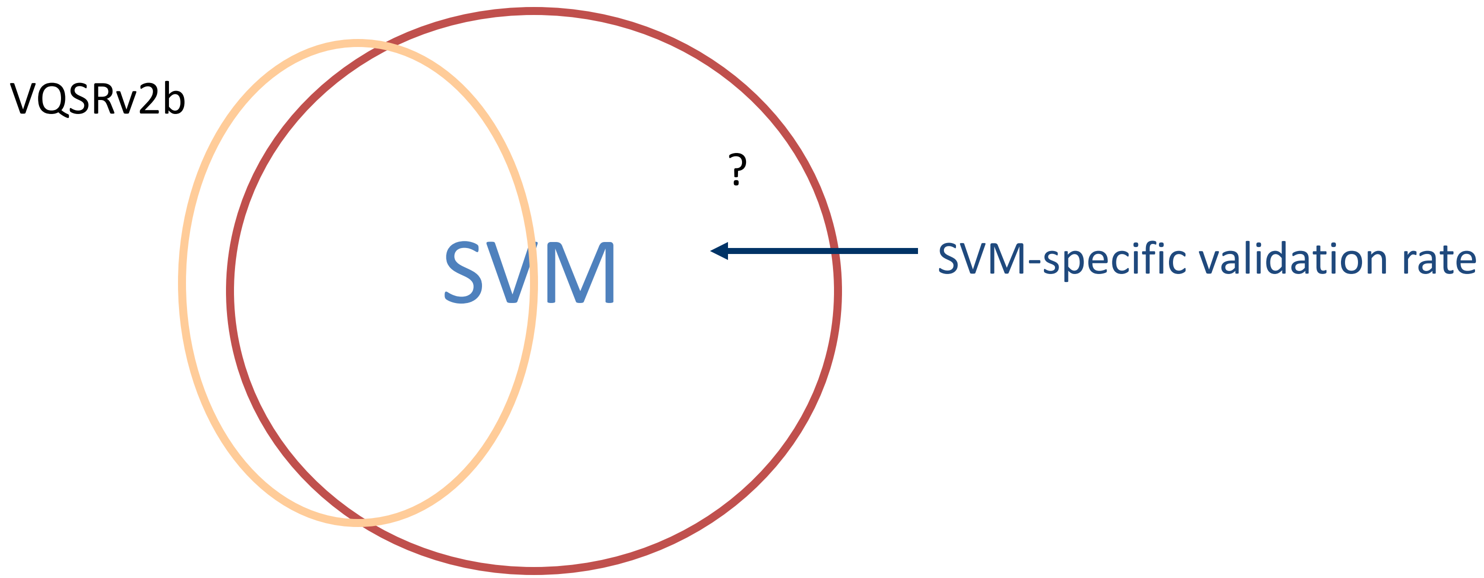


Overall FDR

SVM	# total sites	# sites picked
singleton	5372	100
< 1%	4430	50
>= 1%	1896	50
Total	11698	200

* At most 5 samples for non-singleton

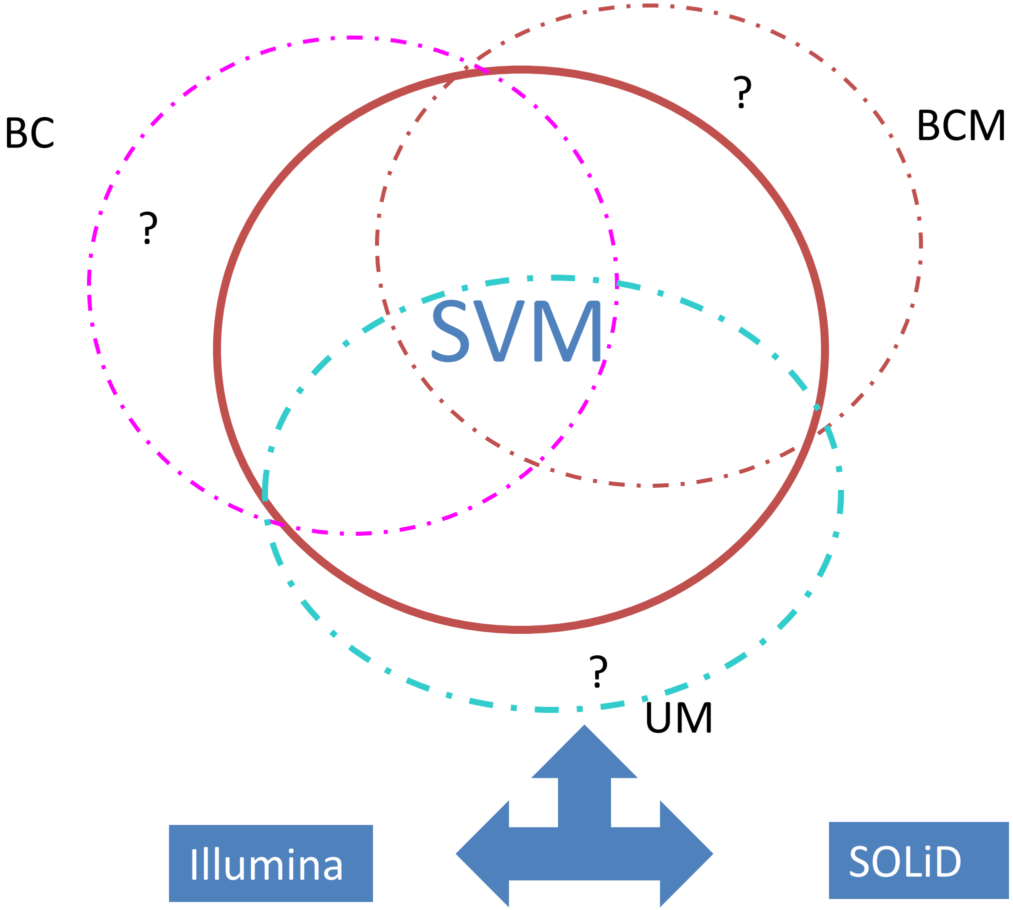
Validation design (2): SVM-specific quality



	#total sites	Unique to VQSRv2b	After excluding latest dbSNP	# sites picked
SVM_all	11698	4887	3327	100

* At most 2 samples for non-singleton

Validation design (3) : Individual callset quality



BC

?

BCM

?

SVM

?

UM

Illumina

SOLiD

Centers	Platform	Total	Outside of SVM	After excluding latest dbSNP and VQSR v2b	# sites picked
BC	ILLUMINA	102	74	54	20
BCM	ILLUMINA	234	157	110	20
UM	ILLUMINA	249	74	43	20
BC	SOLID	91	28	17	17
BCM	SOLID	438	200	155	20
UM	SOLID	200	117	110	20

Summary of 454-PCR validation

- In total, 417 sites of 419 samples are picked. The number of events is 834.
- Samples of different sites are amplified and pooled together for 454 sequencing
- Assign different pools for samples of the same site

Results of (1) SVM overall and (2) SVM specific

	total	submitted	yield	validated	validated/yield
singleton	5372	100	93	92	98.9%
<1%	4430	50	49	47	95.9%
>1%	1896	50	46	46	100%
SVM overall	11698	200	188	185	98.4%
SVM exclusive to VQSRv2b and dbSNP	3327	100	86	84	97.7%

- In total, 269 out of 274 yielded exome consensus sites are validated (98.2%)
- only pick a subset of samples in validation, so the lower limits are measured

Diagnosis of the “failed” sites

Chr	Pos	Site source	AC	Sample	PCR-454 validation	Integrated genotype release	SNPtools	BBMM integrated GL in log-10 scale RR/RA/AA	Exome calls (BCM)
20	20033172	EX_SOLID	singleton	NA19468 (SOLID)	0/0	0/1	0/1	./.:-5,-0.000391054,-3.04576	0/1
20	23667835	EX_ILLUMINA	<1%	NA18510 (Illumina)	0/0	0/1	0/1	./.:-5,-0.00020851,-3.31876	0/0 or ./.
20	23667835	EX_ILLUMINA	<1%	NA18858 (Illumina)	0/0	0/1	0/1	./.:-2.72124,-0.000825952,-5	0/0 or ./.
20	25478962	EX_ILLUMINA	<1%	HG00104 (SOLiD)	0/0	0/1	0/1	./.:-5,0,-5	0/0 or ./.
20	25478962	EX_ILLUMINA	<1%	HG00234 (SOLiD)	0/0	0/1	0/0	./.:-3.1549,-0.000304111,-5	0/0 or ./.
20	25478962	EX_ILLUMINA	<1%	HG00364 (SOLiD)	0/0	0/1	0/1	./.:-4.69838,-8.69777e-06,-5	0/0 or ./.
20	25478962	EX_ILLUMINA	<1%	HG00593 (SOLiD)	0/0	0/1	0/0	./.:-3.1938,-0.000278053,-5	0/0 or ./.
20	25478962	EX_ILLUMINA	<1%	HG01271 (SOLiD)	0/0	0/1	0/0	./.:-0.31142,-0.290883,-5	0/0 or ./.
20	60885811	EX_ILLUMINA	<1%	HG00134 (SOLiD)	0/0	0/1	0/0	./.:-0.477139,-0.477113,-0.477113	0/0 or ./.
20	60885811	EX_ILLUMINA	<1%	HG00350 (SOLiD)	0/0	0/1	0/0	./.:-0.123447,-0.61343,-2.41117	0/0 or ./.
20	62326235	EX_ILLUMINA	<1%	HG00128 (SOLiD)	0/0	0/1	0/0	./.:-4.22169,-2.6068e-05,-5	0/0 or ./.
20	62326235	EX_ILLUMINA	<1%	HG00179 (SOLiD)	0/0	0/1	0/0	./.:-3.22182,-0.000773747,-2.92812	0/0 or ./.

- Most failed events are not called in exome’s individual calls
- SNPtools and Begale disagree with each other half and half on the failed events
- The failed events are probably imputation artifacts
- The sites may still be SNP, but not lucky to pick the “right” samples in the validation

Genotypes concordance on consensus sites

454 PCR validation

Beagle

	0/0	0/1	1/1
0/0	NA	29	NA
0/1	7	530	2
1/1	NA	2	37

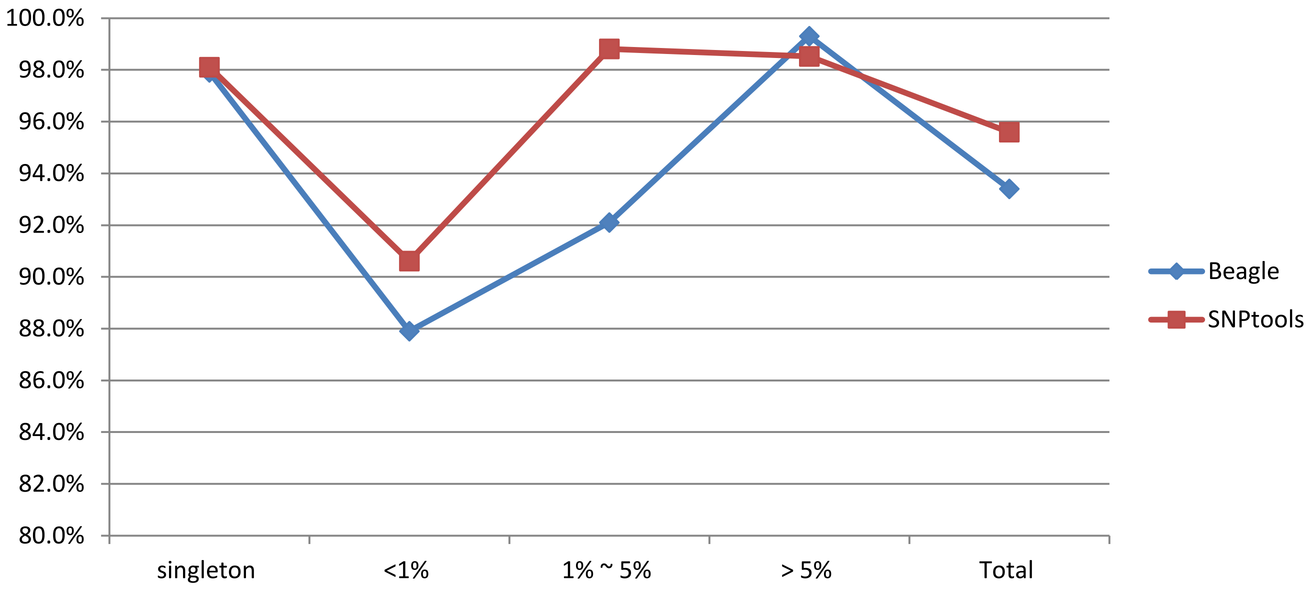
Genotype Concordance = 93.4%

SNPtools

	0/0	0/1	1/1
0/0	13	13	NA
0/1	8	514	3
1/1	NA	2	37

Genotype Concordance = 95.6%

Genotypes concordance by AC



Beagle

AF	singleton	<1%	1% ~ 5%	> 5%	Total
yield	142	240	89	136	607
validated	139	211	82	135	567
validated/yield	97.9%	87.9%	92.1%	99.3%	93.4%

SNPtools

AF	singleton	<1%	1% ~ 5%	> 5%	Total
yield	158	213	84	135	590
validated	155	193	83	133	564
validated/yield	98.1%	90.6%	98.8%	98.5%	95.6%

Results of (3) individual call set quality

	total	selected by SVM	outside of SVM/VQSR/ dbSNP	submitted	yield	validated	validated/yield
Illumina							
BC specific	102	28	54	20	6	2	33.3%
BCM specific	234	77	110	20	18	13	72.2%
UM specific	249	175	43	20	14	13	92.9%
SOLiD							
BC specific	91	63	17	17	13	6	46.2%
BCM specific	438	238	155	20	18	4	22.2%
UM specific	200	83	110	20	16	0	0.0%

- UM's Illumina calls is impressive
- In total, 28 out of 38 Illumina sites are validated (73.7%), 10 out of 47 SOLiD sites are validated (21.3%)

Conclusion

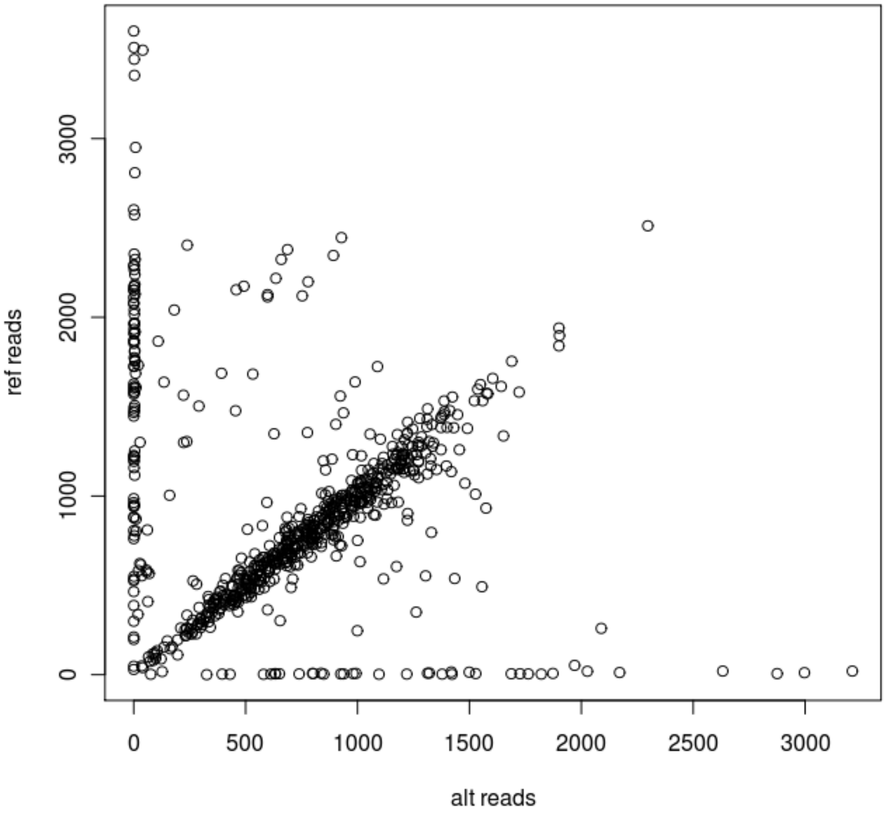
- Exome consensus sites are of high quality
  - 98.9% validation rate for singleton
  - 97.7% validation rate for novel sites
  - 98.2% validation rate for all the sites
  - These are lower estimation due to imputation artifacts and limited capacity to choose samples
- Overall genotype concordance is 93.4% for Beagle and 95.6% for SNPtools, most imputation errors happen in low frequency bins
- Exome SVM consensus strategy is conservative
  - 73.4% unselected Illumina calls are validated as true SNP

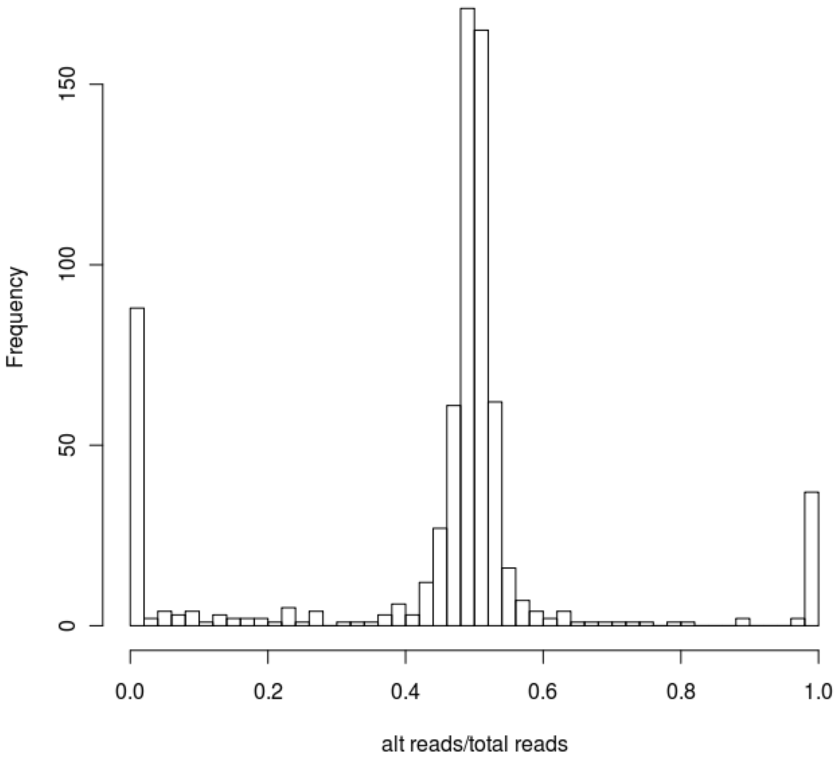
Appendix

Data

- Official integrated genotypes
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20111111_old_phase1_release_files/ALL.chr20.merged_beagle_mach.20101123.snps_indels_svs.genotypes.vcf.gz
- SNPtools integrated genotypes
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20111007_bcm_intergrated_genotypes/All.chr20.LC1041_E1041_Integrated_GT.20101123_20110521.snp.genotypes.vcf.gz
- BBMM GL
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110826_genotype_likelihoods/snps/All.chr20.LC1041_E1041_UNION_GL.20101123_20110521.snp.lc_and_exome.genotypes.vcf.gz
- Exome individual call sets
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110721_exome_call_sets/
  - ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110926_exome_call_sets/ALL.BCM_SOLiD_Bfast_ontarget_plus50bpflanks_306_v3.20110521.snp.exome.genotypes.vcf.gz

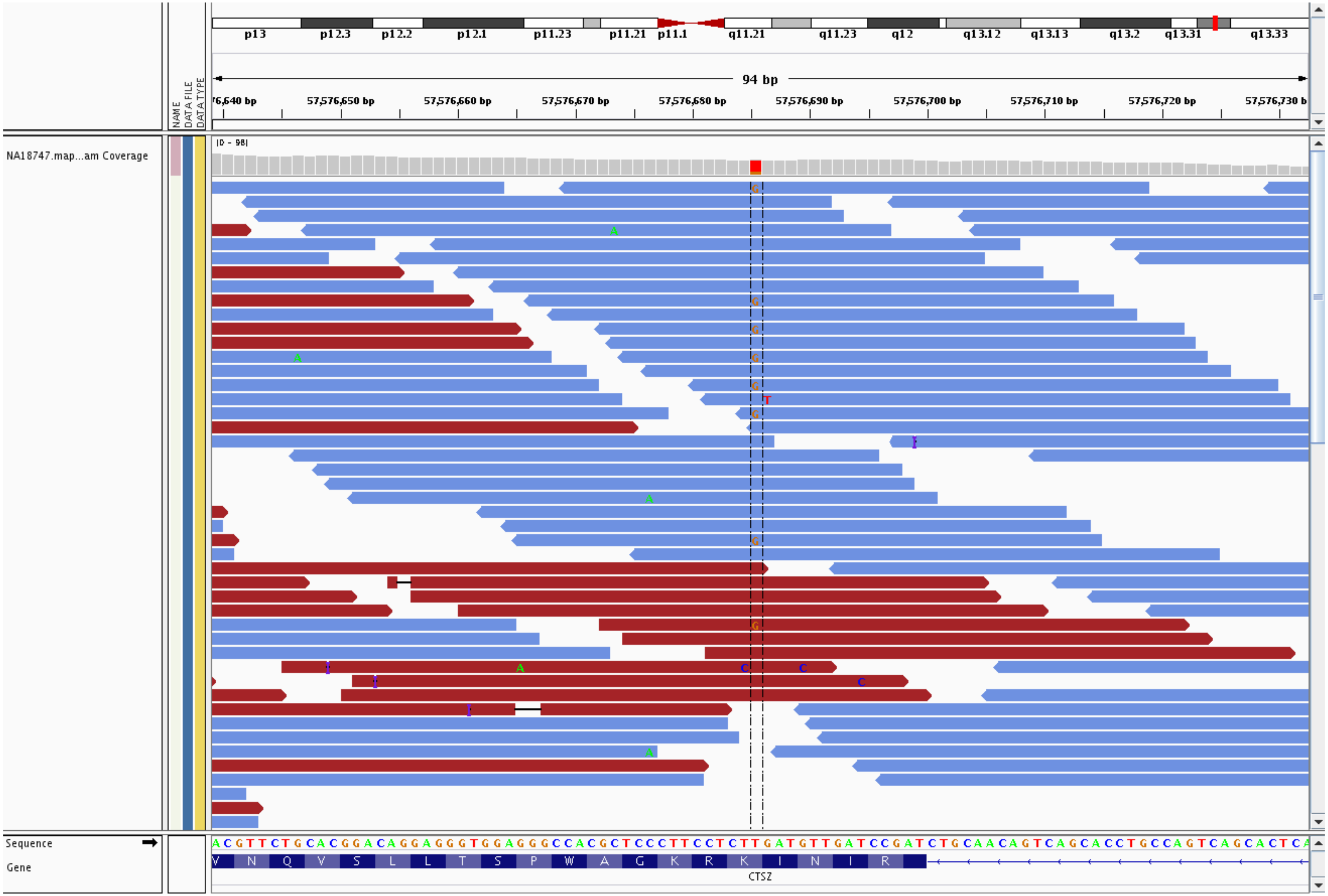
Raw 454-PCR validation data





- Signals are strong in general
- Two swapped samples were withdrew in QC

An example of FP SOLiD calls



Non random errors on both strands