Improving Specificity in Ion Proton Data

Daniel S. Lieber[1], Niru Chennagiri[1], Eric White[1], Dumitru Brinza[2], Jim Veitch[2], Timothy Yu[1], John F. Thompson[1]
1) Claritas Genomics, Cambridge, MA; 2) Thermo Fisher Scientific, San Francisco, CA

CLARITAS GENOMICS

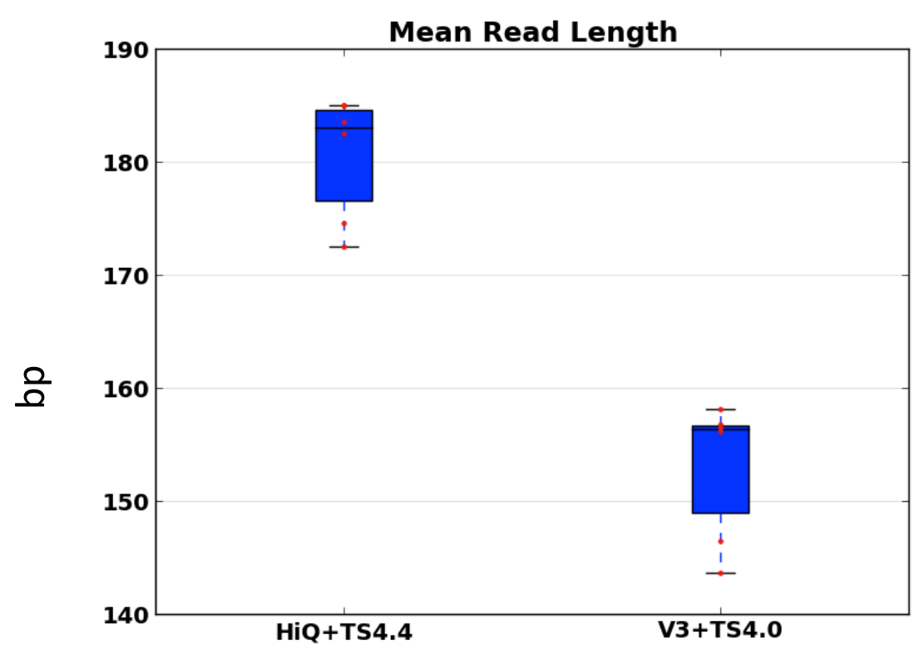
ABSTRACT

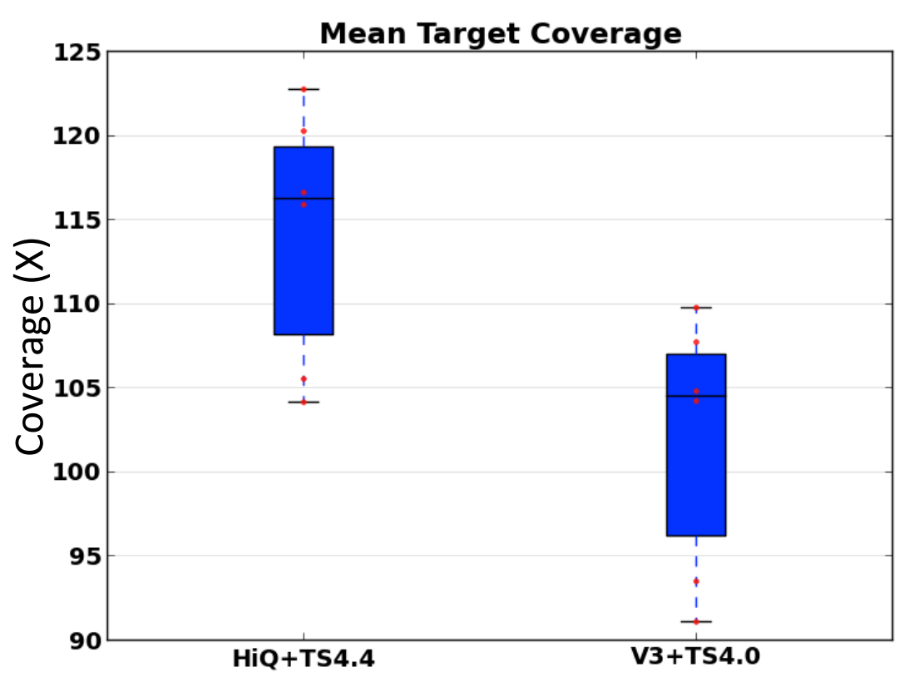
The Ion Proton instrument by Life Technologies is a chip-based sequencing technology that can produce whole-exome scale next generation sequencing data within hours. A critical part in development of a sequencing platform involves improving specificity via data filtering and thus we have worked to improve Ion Proton variant calls using a dual strategy. We first created and applied a strand-specific error (SSE) filter, whereby variants with strand bias across multiple samples were removed. We additionally created a filter using data derived from over 8,000 exomes sequenced at Claritas Genomics and compared the allele frequencies to the ExAc database. Any variant with high allele frequency in the Claritas database but low allele frequency in ExAc was deemed to be a likely false positive. Using NA12878 NIST data to benchmark performance, these methods allowed us to reduce the number of false positive SNPs by 3.8-fold and indels by 5.6-fold, with a 0.1% loss in sensitivity for both SNPs and indels. These results demonstrate that Ion Proton variant call accuracy can be significantly improved through downstream bioinformatics processing.

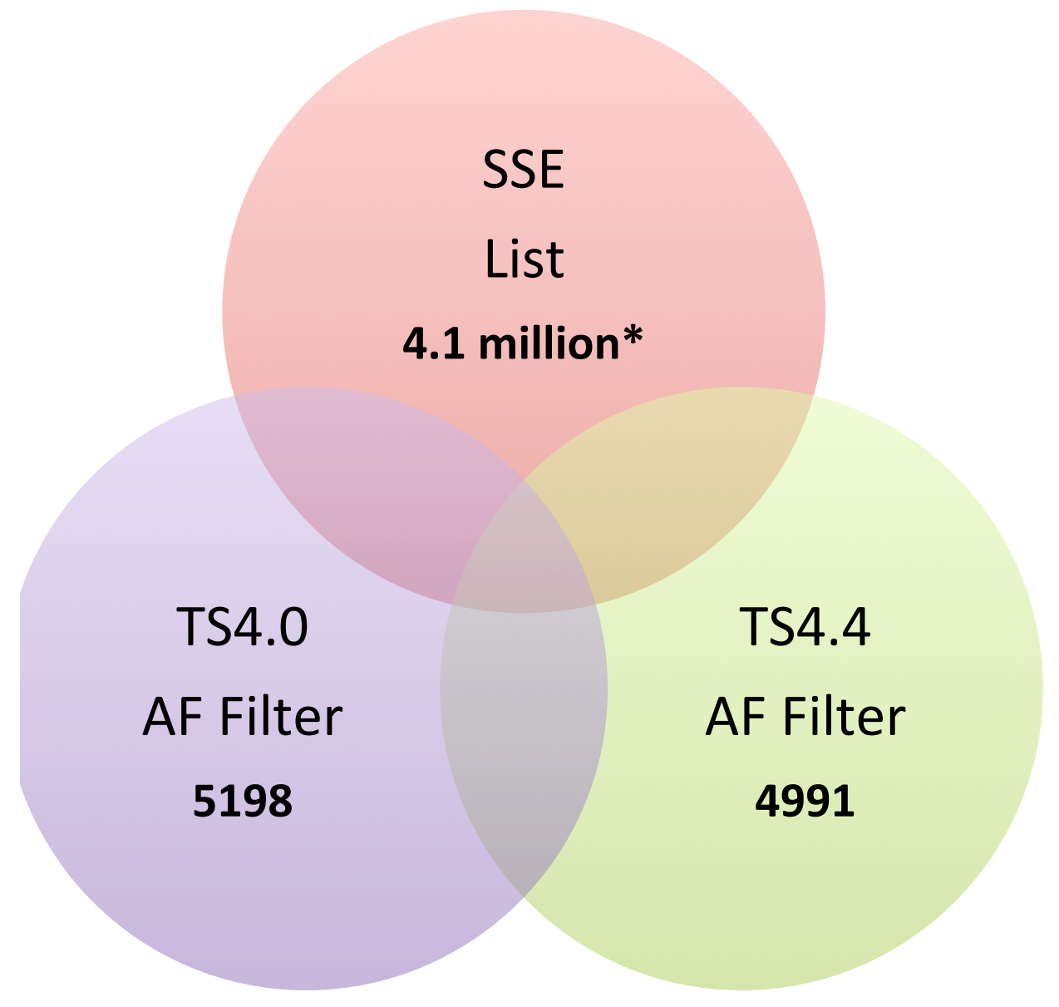
Methods

Name	Chemistry	Software	# Exomes
TS4.0	V3	Torrent Suite 4.0	8,003
TS4.4	HiQ	Torrent Suite 4.4	4,479

- Analysis region: AmpliSeq.intersected.nist.intersected.ExAC (~N Mb)
- NIST version: ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.17/
- Sensitivity / Specificity (Sn/Sp): Calculated by comparing NA12878 exome data to NIST Genome In A Bottle (GIAB), using proprietary software on the analysis region specified above. Zygosity differences are considered non-matches.
- Sn/Sp results: NA12878 HiQ TS4.4 data (mean values, N=2)
- Allele Frequency Filter: List of false positive positions defined as variants that are seen at >1% allele frequency in Ion Proton data but at <0.1% allele frequency in ExAC (N=60,706).
- Strand Specific Error Filter: Based on 12 AmpliSeq TS4.0 exomes. Generated by ThermoFisher using a proprietary algorithm. Identifies likely strand-specific errors. Filters variants that are both on the strand bias blacklist and show a similar strand bias in the analyzed sample.

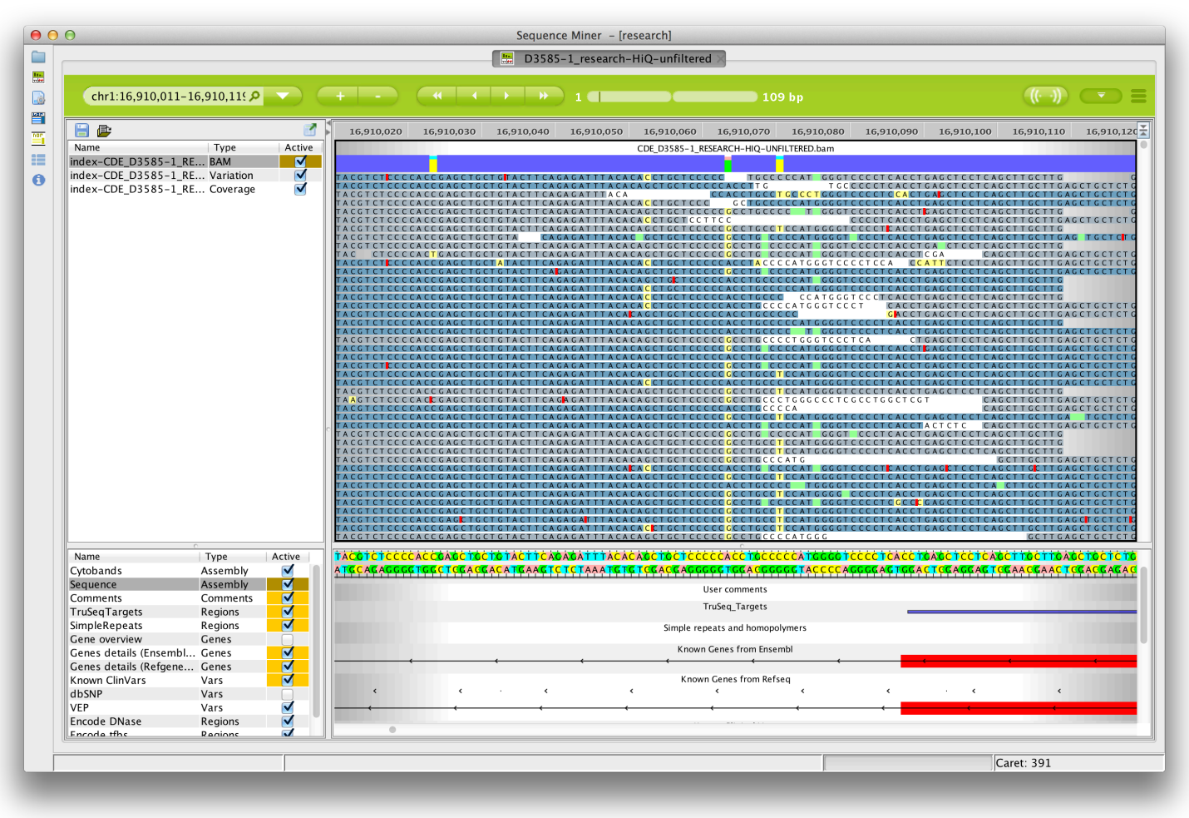


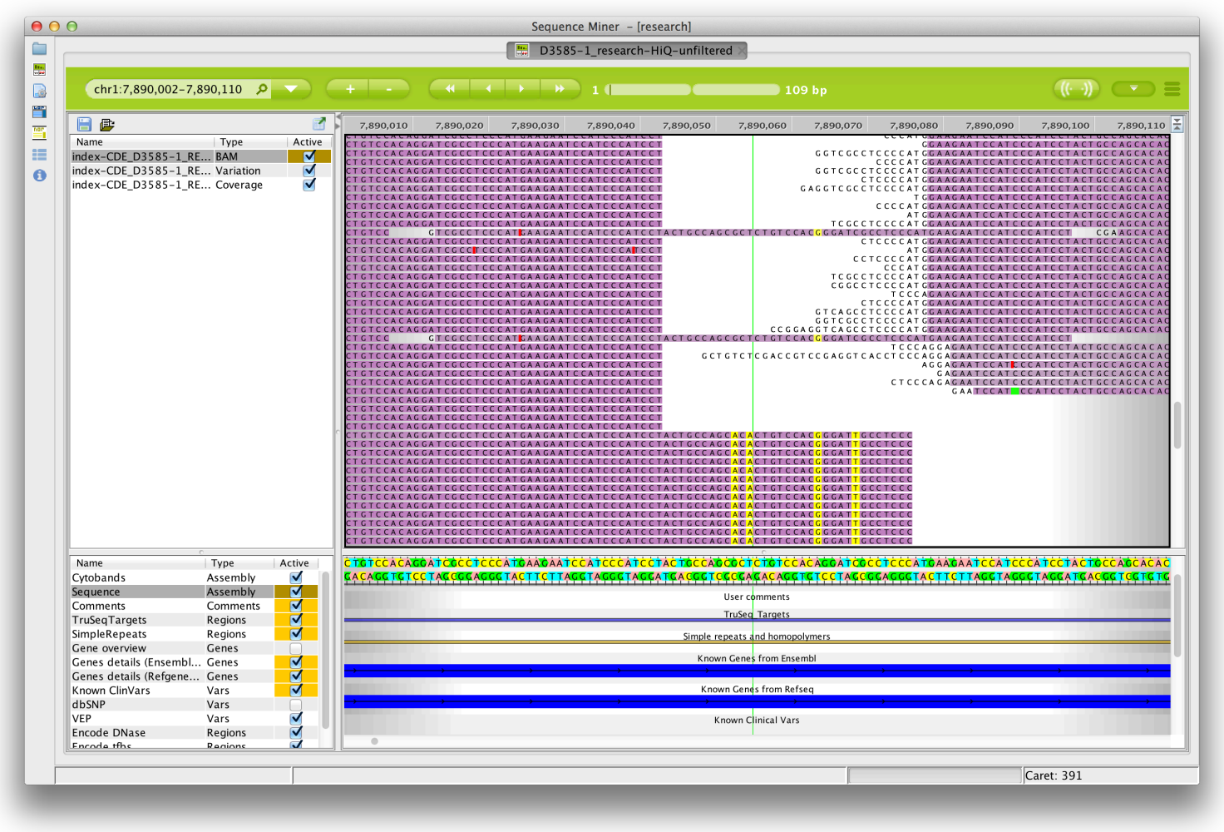




Results

	Sensitivity	Specificity (FPs/Mb)	PPV
TS4.0 Unfiltered			
SNP	88.3%	6.2	99.1%
INDEL	45.6%	12.5	61.3%
TS4.0 AF Filtered			
SNP	88.3%	2.0	99.7%
INDEL	45.6%	4.2	82.3%
TS4.0 SSE Filtered			
SNP	88.3%	4.7	99.3%
INDEL	45.6%	7.6	72.1%
TS4.0 AF + SSE Filtered			
SNP	88.3%	1.8	99.7%
INDEL	45.6%	2.5	88.8%
TS4.4 Unfiltered			
SNP	96.0%	8.4	98.9%
INDEL	51.4%	2.1	91.4%
TS4.4 AF Filtered			
SNP	96.0%	2.7	99.6%
INDEL	51.4%	0.6	97.2%





TS4.4 AF likely false positives screenshots

Conclusions

- Ion Torrent Proton technology has improvied significantly in last several years
- HiQ chemistry + TS4.0 >> v3 chemistry + TS4.0
- Computational methods enable one to reduce the number of false positives by >3-fold with a minimal loss of sensitivity.

References

1. Rothberg, J.M. et al, An integrated semiconductor device enabling non-optical genome sequencing. Nature. (2011)
2. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. 32, 246–51 (2014).
3. Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org) [28 Sept 2015].

See other Claritas Genomics posters:
1933, 1981, 2070, 2071, 2085

CLARITAS GENOMICS INC.
99 Erie Street | Cambridge, MA 02139

clientservices@claritasgenomics.com
p: 617.553.5880 p: 855.373.9003 f: 617.553.5842