The Sentieon Genomic Tools
Improved Best Practices Pipelines for Analysis of Germline and Tumor-Normal Samples

Andreas Scherer, Ph.D. President and CEO
Dr. Donald Freed, Bioinformatics Scientist, Sentieon

20 most promising Biotech Technology Providers
Top 10 Analytics Solution Providers
Hype Cycle for Life sciences
Golden Helix is a global bioinformatics company founded in 1998.

**Variant Calling**
- Filtering and Annotation
- Clinical Reports
- CNV Analysis
- Pipeline: Run Workflows

**Variant Warehouse**
- Centralized Annotations
- Hosted Reports
- Sharing and Integration

**SNP & Variation Suite**
- GWAS
- Genomic Prediction
- Large-N-Population Studies
- RNA-Seq
- Large-N CNV-Analysis
Cited in over 1,100 peer-reviewed publications
Over 350 customers globally
Golden Helix – Who We Are

When you choose a Golden Helix solution, you get more than just software

- REPUTATION
- TRUST
- EXPERIENCE
- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS
- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS
End-to-End Architecture for Clinical Testing Labs

PRODUCTS
- VS-CNV
- SENTIEON DNAseq
- SENTIEON TNseq

BIOINFORMATICS PIPELINE
- FASTQ
- BAM
- VCF

ANNOTATED VCF
- OMIM
- CADD
- SNV & POLYPHEN
- SIFT & POLYPHEN
- 1000G EXOMES
- CONSERVATION SCORES
- CLINVAR
- DBSMP
- COSMIC
- ENSEMBL GENES
- REFSEQ GENES

FUNCTION
- SINGLE NUCLEOTIDE VARIATION
- COPY NUMBER VARIATION & LOSS OF HETEROZYGOITY
- CHROMOSOMAL ABERRATION
- PUBLIC & COMMERCIAL ANNOTATIONS TO ENRICH GENOMIC DATA SETS
- ANNOTATE & FILTER VISUALLY INSPECT ALIGNMENTS VARIANT PRIORITIZATION CLINICAL ASSESSMENT

CLINICAL REPORT
- VARSEQ
- VARSEQ REPORTS
- VARSEQ PIPELINE

DATA WAREHOUSING
- WAREHOUSE
- WEB-ENABLED INTERFACE
+ POWERFUL API: JSON, XML
TSV, CSV, SQL, FHIR

CLINICAL ASSESSMENT CATALOG
ADVANCED DATA QUERYING VERSIONING
INTEROPERABILITY COMPLIANCE WITH HIPAA, CLIA, & CAP DATA DISCOVERY
The Sentieon Genomic Tools – Enabling Precision Data for Precision Medicine

WWW.SENTIEON.COM
What is Precision Data

• Hottest Word: Big Data
• But, PRECISION is the goal:
  precision recommendation, precision prediction
• Big genomics data for precision medicine:
  -precision diagnostics
  -improve precision treatment for individual

Big + Accurate  |  “better data in, better results out”

→ Precision Data
Sentieon’s Mission

Enable Precision Genomics Data for Precision Medicine

- Ability to process big data at affordable cost and time
- With confidence
  - The highest accuracy
  - Consistent results
Three components of analytical software

- **Mathematical methods**
- **Same mathematical models as the Broad Institute**

- **Compute algorithms**
- **More efficient compute algorithms**

- **Software implementation**
- **Enterprise strength software implementation**

---

**Diagram**: A flowchart showing the interaction between different parts of the software components. Each part is connected with arrows indicating the flow of data or processes.

---

**Links**:
- [Three components of analytical software](http://www.cs.princeton.edu/courses/archive/spr05/cos226/assignments/baseball/)
<table>
<thead>
<tr>
<th>Mission</th>
<th>Products</th>
<th>Awards</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNAseq</td>
<td>DNAseq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNseq</td>
<td>TNseq</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DNAseq**
- Identical* results as Broad Institute’s “Best Practice Workflow”
- BWA-GATK HaplotypeCaller

**TNseq**
- Identical* results as Broad Institute’s “Somatic Variant Discovery Workflow”
- MuTect and MuTect2

- Identical math, much more efficient computing algorithm and enterprise-strength software engineering
- 10+ X faster whole pipeline FASTQ to VCF; 20X-50X faster GATK/MuTect/MuTect2 portion; in core-hours
- Both products kept up-to-date with Broad Institute’s releases

*1/1000 vcf differences due to GATK down-sampling, thread dependency, rounding differences
GATK is the gold standard

GATK Haplotype Caller is the most accurate DNA analysis tool

- “As of GATK version 3.3, we recommend using HaplotypeCaller in all cases, with no exceptions.”
  - Broad Institute

- “Haplotype Caller is more accurate than the Unified Genotyper”
  - https://hpc.mssm.edu/files/Carneiro_workshop.pdf

- “GATK HaplotypeCaller called a substantial number of indels not called using VarScan-Cons (as well as GATK UnifiedGenotyper)”
GATK is the gold standard

GATK Haplotype Caller is the most accurate DNA analysis tool

But GATK Haplotype Caller is too slow

Other speedup efforts:
(1) massive parallel on cloud (challenges: workflow, data privacy, cost) – may combine with our solution;
(2) hardware acceleration (challenges: cost, intrusive, inflexibility, scalability);
(3) corner-cutting (challenge: lose info)

Our approach: stay with the most accurate math, but use much more efficient compute algorithm with enterprise-strength software implementation
**Mission**

**Products**

- DNAseq
- TNseq

**Awards**

**Value**

---

*Server specs: 32 core 2.4 GHz Intel Xenon server, 64 GB memory*
DNAseq  

Highlights beyond speed:

• 100% consistency, no run-to-run differences
  No down-sampling in high coverage region, no thread dependency
  ➔ Higher accuracy by eliminating software noise

• System robustness
  ➔ Large dataset joint call over 100K samples together
    (without intermediate merging)
Consistency Challenge
(2/25-4/25/2016)
- Top Overall Performance
- Highest Reproducibility

Truth Challenge
(4/26-5/26/2016)
- Highest INDEL Precision
- Highest SNP Recall

*Screenshots from https://precision.fda.gov/
ICGC-TCGA DREAM Mutation Calling challenge

Annual open contest by ICGC-TCGA for somatic variant calling accuracy.

Challenge-6 due 8/19/2016 (extended from 4/22/2016)
ICGC-TCGA DREAM Mutation Calling challenge

Final Leaderboard (8/19/2016)

<table>
<thead>
<tr>
<th></th>
<th>SNV</th>
<th>INDEL</th>
<th>SV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentieon</td>
<td>98.57%</td>
<td>98.14%</td>
<td>100%</td>
</tr>
<tr>
<td>on 4/21</td>
<td>98.17%</td>
<td>on 4/21 97.48%</td>
<td>on 4/21 100%</td>
</tr>
<tr>
<td>Bina/Roche</td>
<td>97.57%</td>
<td>Bina/Roche 97.01%</td>
<td>Genowis 99.82%</td>
</tr>
<tr>
<td>Genowis</td>
<td>96.92%</td>
<td>OICR-GSI 86.99%</td>
<td>Gridss 99.63%</td>
</tr>
</tbody>
</table>

-Sentieon leads in all categories-
**Consistency**

precisionFDA Consistency Challenge Reproducibility

F1-score(%) between runs and between samples

<table>
<thead>
<tr>
<th>Garvan vs. Garvan rerun</th>
<th>Garvan vs HLI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FP</strong></td>
<td><strong>FP</strong></td>
</tr>
<tr>
<td>Sentieon by UNM*</td>
<td>0</td>
</tr>
<tr>
<td>Sentieon (dual mapping)</td>
<td>0</td>
</tr>
<tr>
<td>ISAAC**</td>
<td>0</td>
</tr>
<tr>
<td>Genalice**</td>
<td>3621</td>
</tr>
<tr>
<td>Edico Dragen**</td>
<td>3147</td>
</tr>
<tr>
<td>107302</td>
<td>295568</td>
</tr>
<tr>
<td>112255</td>
<td>266952</td>
</tr>
<tr>
<td>286306</td>
<td>433504</td>
</tr>
<tr>
<td>161213</td>
<td>315611</td>
</tr>
</tbody>
</table>

* Sentieon standard pipeline results from Jeremy Edwards (University of New Mexico) submission
** Edico, Genalice (MAP 2.2.0) and Isaac (aligner v01.14.07.14 & variant caller v2.0.13) run results from Changhoon Kim's (Macrogen Clinical Laboratory) submission
### Accuracy

#### precisionFDA Consistency Challenge Accuracy

**F1-score(%) to NIST truth set**

<table>
<thead>
<tr>
<th></th>
<th>Garvan sample</th>
<th>HLI sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>SNP only</td>
</tr>
<tr>
<td>Sentieon by UNM*</td>
<td>99.39</td>
<td>99.86</td>
</tr>
<tr>
<td>Sentieon (dual mapping)</td>
<td>99.47</td>
<td>99.88</td>
</tr>
<tr>
<td>ISAAC**</td>
<td>97.29</td>
<td>98.57</td>
</tr>
<tr>
<td>Genalice**</td>
<td>98.04</td>
<td>99.20</td>
</tr>
<tr>
<td>Edico Dragen**</td>
<td>99.25</td>
<td>99.74</td>
</tr>
</tbody>
</table>

* Sentieon standard pipeline results from Jeremy Edwards (University of New Mexico) submission

** Edico, Genalice (MAP 2.2.0) and Isaac (aligner v01.14.07.14 & variant caller v2.0.13) run results from Changhoon Kim’s (Macrogen Clinical Laboratory) submission
Observations:

- Much higher F1 score (~99.9%) than Consistency Challenge, due to masking out complex regions in truth set
- Sentieon has excellent consistency between HG001 and HG002, due to no run-to-run differences
Value of Sentieon solutions

- Highest accuracy: most rigorous math, no noise in algorithm and software
- No down-sampling in high-coverage regions ← Critical for clinical samples
- No run-to-run difference ← Critical for medical decision
  --- proven in precisionFDA challenges and DREAM challenge
- Fast turnaround, ability to scale below 20 minute turnaround → Improved productivity, faster medical decision
- >10X reduced core-hours → drastically reduced compute cost
- Enable large dataset joint call → Enable genomics big data analysis

Sentieon tools: fastest, most accurate, zero run-to-run difference, no down-sampling, large cohort joint call, pure software solution, running on any generic computer systems
- Deployed at >100 sites worldwide (academia and industry)
- Accumulated usage at customer sites:
  >100K WGS/WES, >5e15 bases

Sentieon products are built on the solid foundation of the most rigorous and most extensively validated mathematical models used in Broad Institute’s Best Practice Workflow, but with more efficient computing algorithms and enterprise-strength software implementation.
Contact Don Freed (don.freed@sentieon.com) if you would like to talk to me in person.