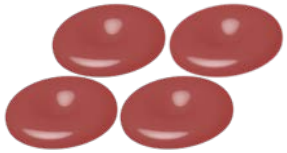
BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Benchmarking of Variant Callers & Analysis Tools

Simon Chan
Assistant Bioinformatics Coordinator
Canada's Michael Smith Genome Sciences Centre
2017-09-27

Illumina Sequencing



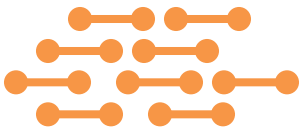
1.) Cells (e.g cancerous or matched normal)

4.) Sequencing

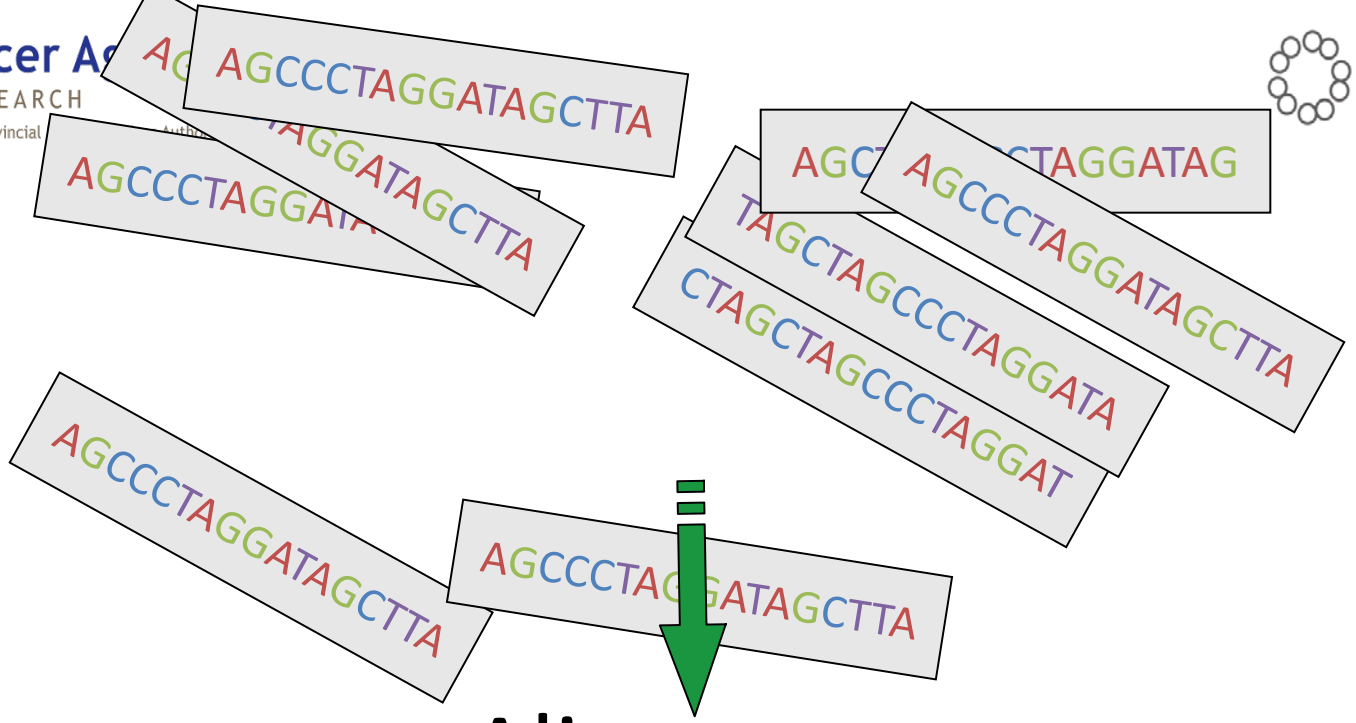
2.) Isolate DNA

5.) Ready for bioinformatics

AAAAAAAAAAAAAAAACCCAAA
CCCTTTTTGGGGAAGGGGGGGTT
TCCCCCCCCCCCCCCCAAAAAAAT
AAAGGGAAAGGGGTTTCCCAAA



3.) Sheared DNA, with sequencing adapters



Alignment

Reference Genome

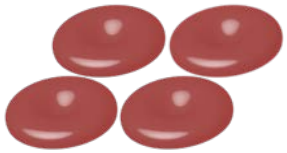
...ACTCGCTAGCTAGCCCTAGGATAGCTTAGAGACCCTCGCGAAATAGACCCTCGAT...

AGCCCTAGGATAGCTTA
AGCTAGCCCTAGGATAG
TAGCTAGCCCTAGGATA
CTAGCTAGCCCTAGGAT
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA

ACCCTCGCGAAATAGAC
ACCCTCGCGAAATAGAC
GACCCTCGCGAAATAGA
AGACCCTCGCGAAATAG
TAGAGACCCTCGCGAAA
CTTAGAGACCCTCGCGA
CTTAGAGACCCTCGCGA

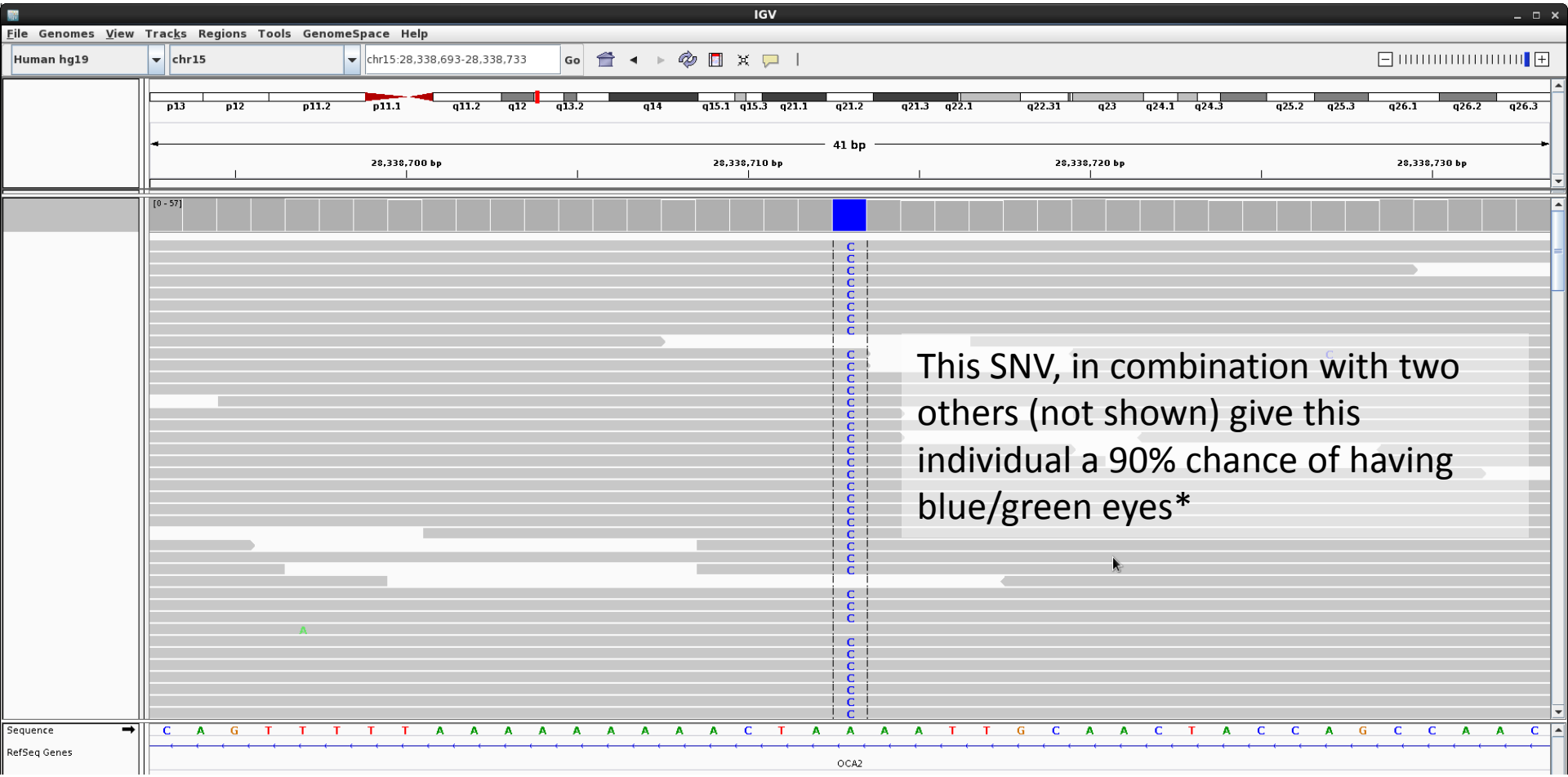
After alignment…

- Binary alignment file (BAM file)
- Binary file reports where in reference genome reads are aligned to
- Cancer BAM vs Matched Normal BAM



1.) Cells (e.g cancerous or matched normal)

Viewing Alignments: Single Nucleotide Variants (SNVs)



*Duffy, David L., et al. "A three–single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation." The American Journal of Human Genetics 80.2 (2007): 241-252.

What are somatic variants?

- Variation in DNA that occurs after conception
- Not in germ cells and thus not passed on to future generations
- Somatic variants may act as cancer drivers
    - e.g. KRAS G12D gain of function mutation in colorectal cancer
- Variant in tumour, not matched normal

Challenges of calling somatic variants

- Sample purity (e.g. biopsy with low tumour content)
- Sequencing biases and errors
- Alignment ambiguities
- Differences in variant calling algorithms

How do we estimate the accuracy of a somatic variant calling pipeline?

Construct a ground truth data set to estimate somatic variant prediction accuracy

Goals:

1. For a cancer sample, **COLLECT** independent somatic variant data sets from different organizations

2. **CURATE** a ground truth set of somatic SNVs and indels

3. **ESTIMATE** accuracy of somatic variants from paired somatic analysis pipeline

COLO-829

- Melanoma cell line
- Isolated from 45 year old Caucasian male
- COLO-829BL is the matched normal made from peripheral blood

Goal #1: Four Independent COLO-829 Somatic Variant Data Sets

Wellcome Trust Sanger Institute

- Pleasance E et al, 2010 (EDP)
- 75 bp reads, tumour/normal: ~40X/~32X
- Sanger validated 497 somatic SNVs and 62 somatic indels

Translational Genomics Research Institute (TGEN)

- Craig DW et al, 2016
- 112 bp reads, tumour/normal: ~80X each

Ground Truth

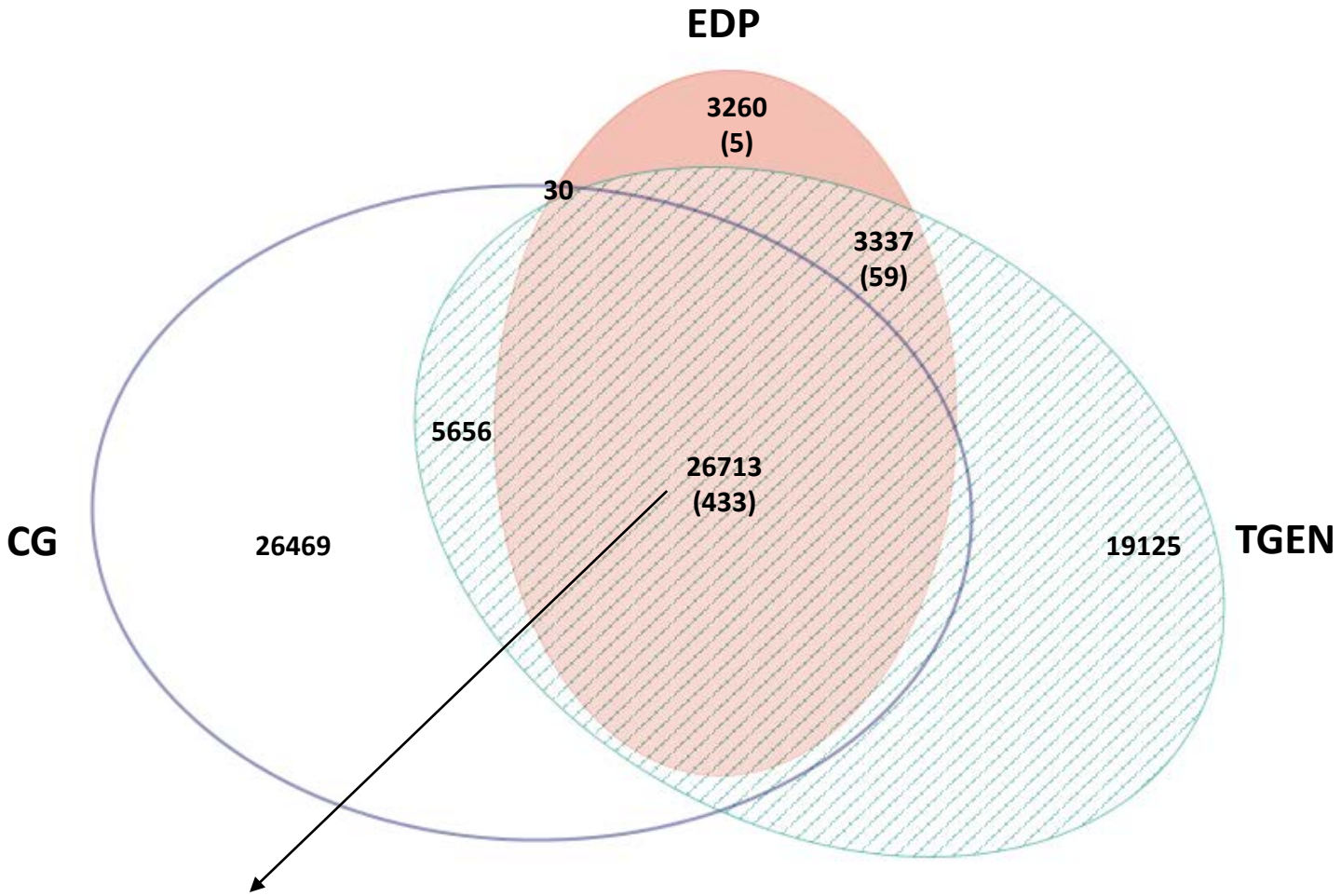
Complete Genomics – BGI (CG)

- Unpublished data
- Proprietary sequencing technology (DNA Nanoball Arrays)

In-house (BCGSC)

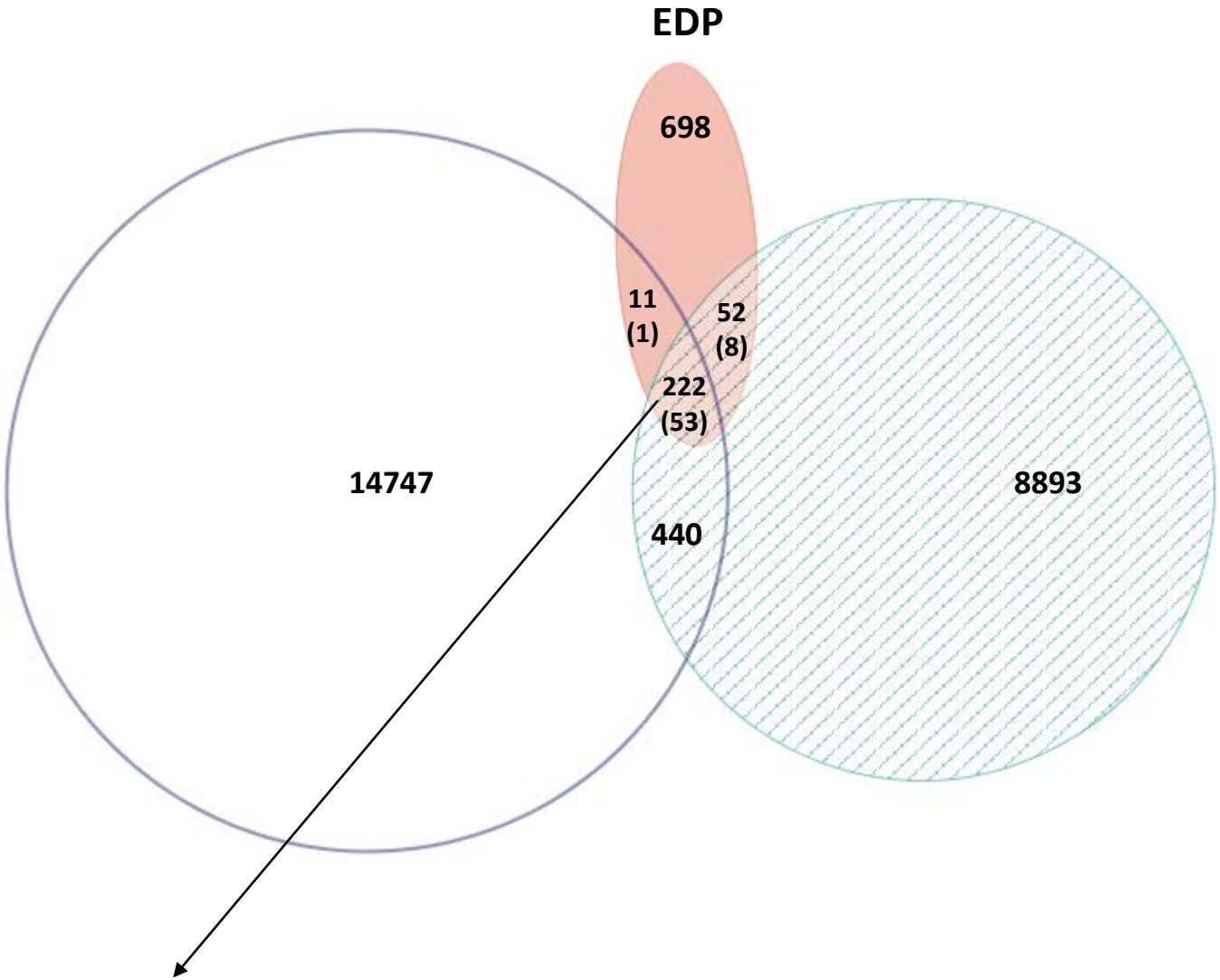
- Craig DW et al, 2016
- 125 bp reads, tumour/normal: ~100X each

COLO-829 Somatic SNV Data Sets



'Tentatively Validated Somatic SNVs'

COLO-829 Somatic Indel Data Sets



'Tentatively Validated Somatic Indels'

Goal #2: Curated ground truth variants

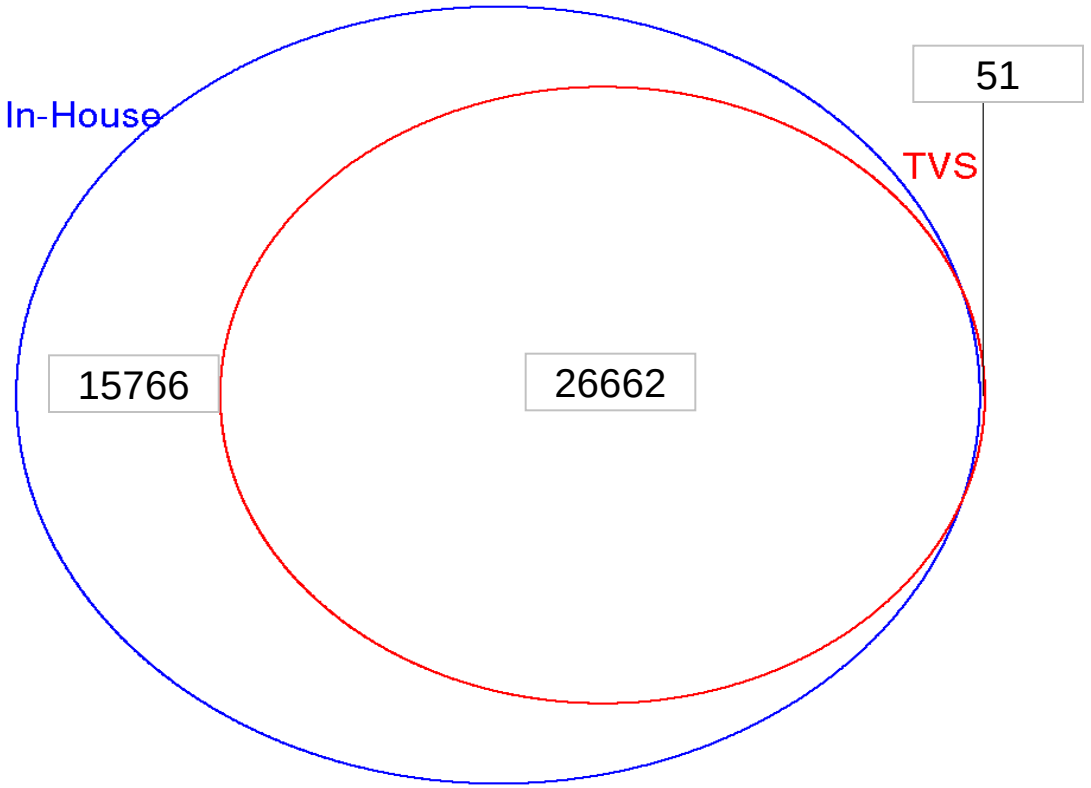
- **Tentatively Validated Somatic SNVs/Indels**
  - These variants **should** be called
  - Any variant that is missed is a potential false negative

- **Union Set of Somatic SNVs/Indels**
  - These represent all possible variants that **could** be called
  - Any **extra** variant is a potential false positive

- **Next step:** Compare our in-house somatic variants to these two data sets.

Somatic SNV Sensitivity

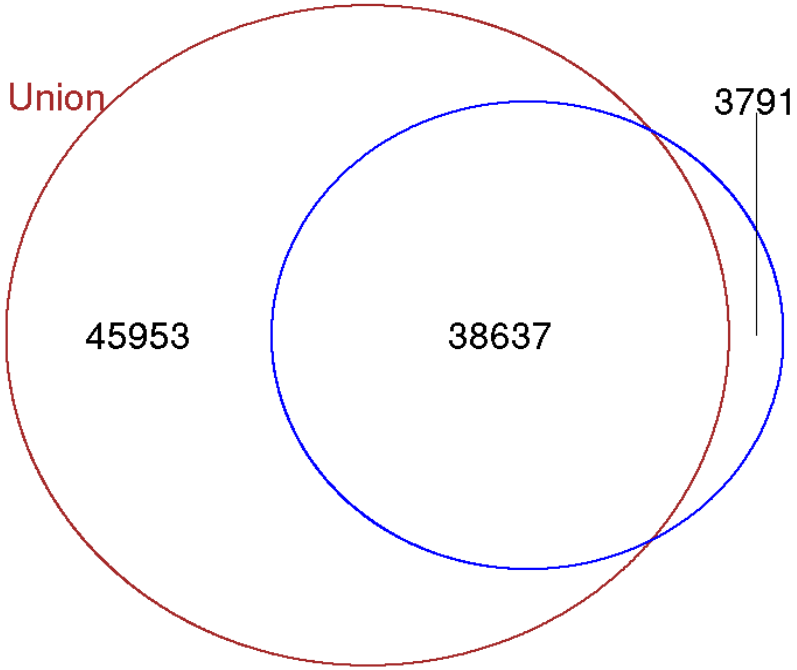
How many of the Tentatively Validated SNVs (TVS, n = 26,713) were called in our In-House Somatic SNVs (n = 42,428)?



Sensitivity Estimate: 26662/26713 ~ 99.8% of TVS were called

Somatic SNV Specificity

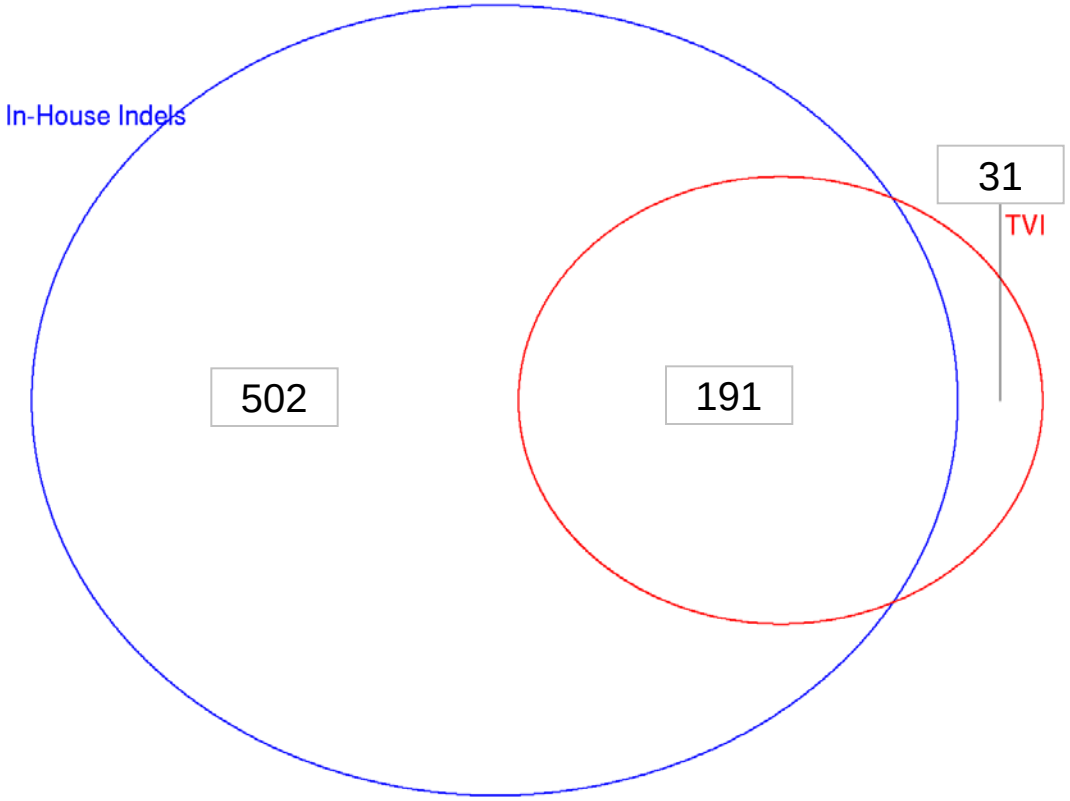
How many of the In-House Somatic SNVs (n = 42,428) were not seen in the **Union SNVs (n = 84,590)**?



Specificity Estimate: 91.1% (3791/42428 ~ 8.9%)

Somatic Indel Sensitivity

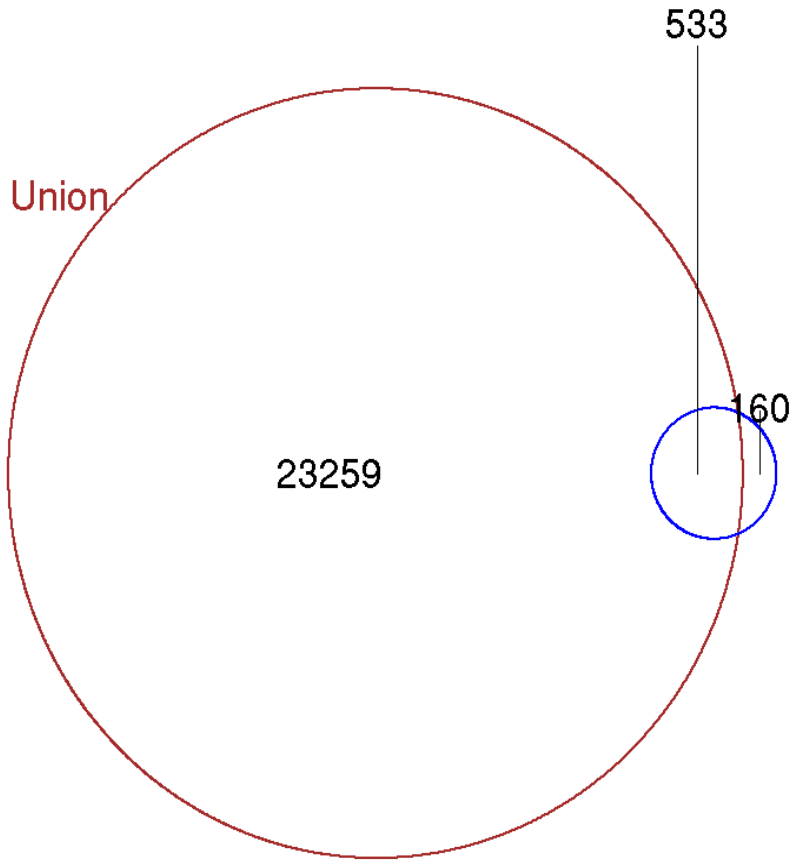
How many of the Tentatively Validated Indels (TVI, n = 222) were called in our In-House Somatic SNVs (n = 693)?



Sensitivity Estimate: 191/222 ~ 86.0% of TVI were called

Somatic Indel Specificity

How many of the In-House Somatic Indels (n = 693) were not seen in the **Union Indels (n = 23,792)**?



Specificity Estimate: 76.9% (160/693 ~ 23.1%)

Review

- Ground truth data set constructed for COLO-829 somatic SNVs and indels and can be used to estimate sensitivity/specificity of somatic predictions

- This data set can be used to benchmark publically available somatic SNV/indel prediction algorithms

SNV Prediction Callers

- Queried BioStars/PubMed/Google for somatic callers
  - Identified 28 tools

- Requirements – Tool is…
  - published in a peer reviewed journal
  - maintained (e.g. bug fixes, updates, etc)
  - supported (e.g. authors respond to questions on BioStarts, etc)
  - outputs Variant Call Format (VCF) file

- 11/28 tools passed these requirements

Benchmarking Tools

- Strategy to evaluate candidate tools for inclusion in production pipelines:
    - Literature search
    - Local installation of tools
    - Construct ground truth set
    - Compare results with respect to evaluation criteria
    - Choose tool(s) to use

- Once every few years, **revaluate** and, if necessary, **update** tools/version

Software	Year Published	Organization
LoFreq 2.1.1	2012	Genome Institute of Singapore
MuTect 1.1.4	2013	Broad
Shimmer 20150410	2013	NHGRI/NIH
FreeBayes 0.9.21	2012	Boston College
Platypus 20150421	2014	The Wellcome Trust Centre for Human Genetics
SAMTools 1.2	2009	Sanger/Broad

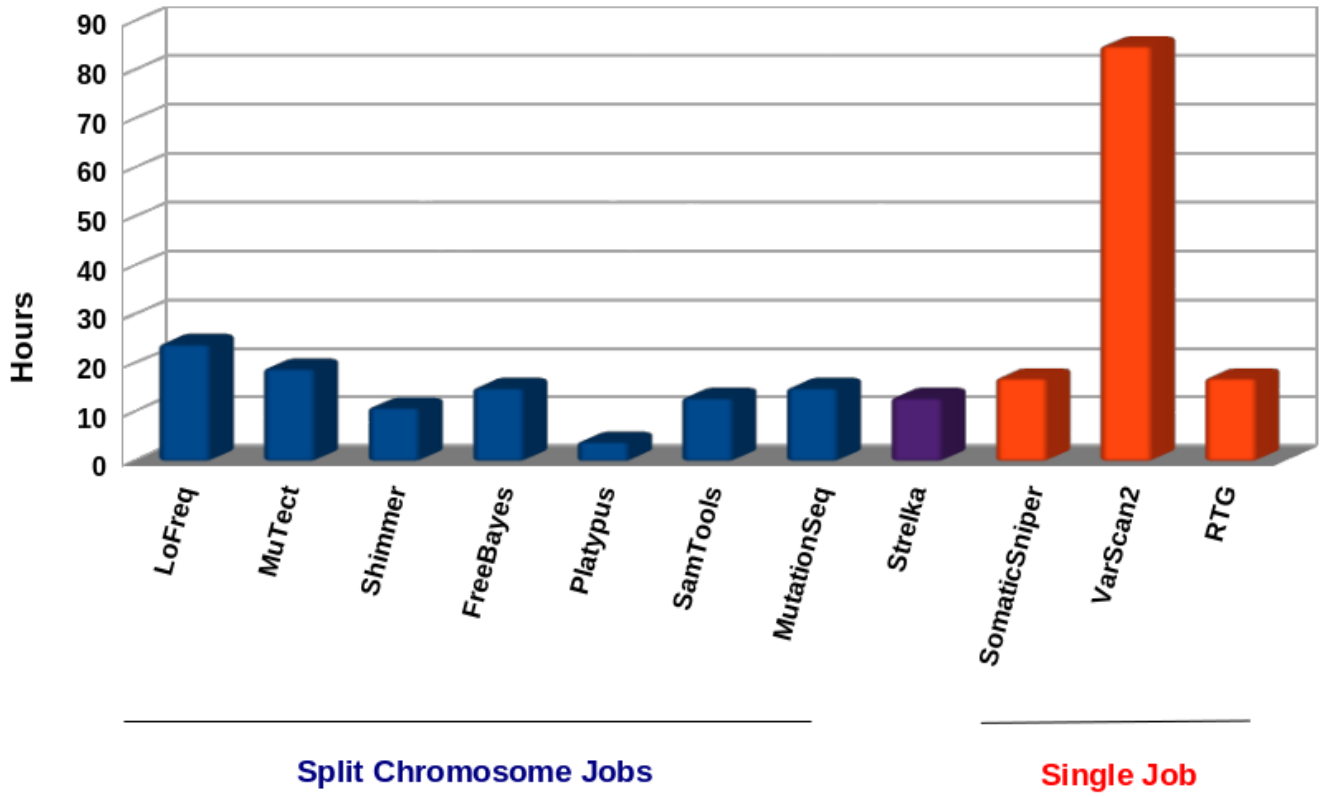
Software	Year Published	Organization
SomaticSniper 20150411	2011	WUSTL
VarScan2 2.3.7	2012	WUSTL
RTG Somatic 3.4.3	2015	Real Time Genomics, Inc
Strelka 1.0.14	2012	Illumina, Inc
MutationSeq 4.3.5	2011	BC Cancer Research Centre

Run each somatic caller on our in-house tumor/normal COLO-829 sample

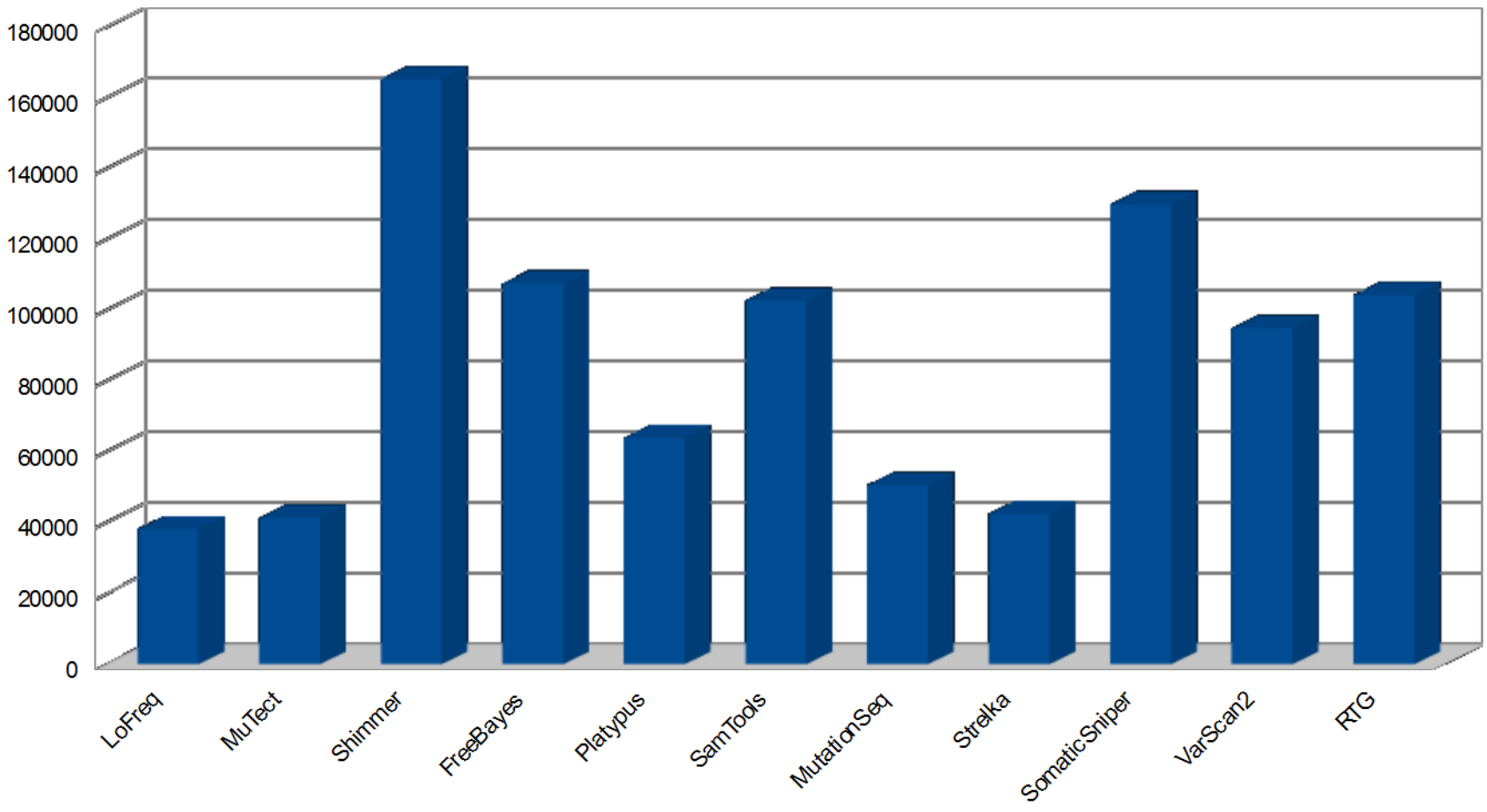
For each somatic caller, report:

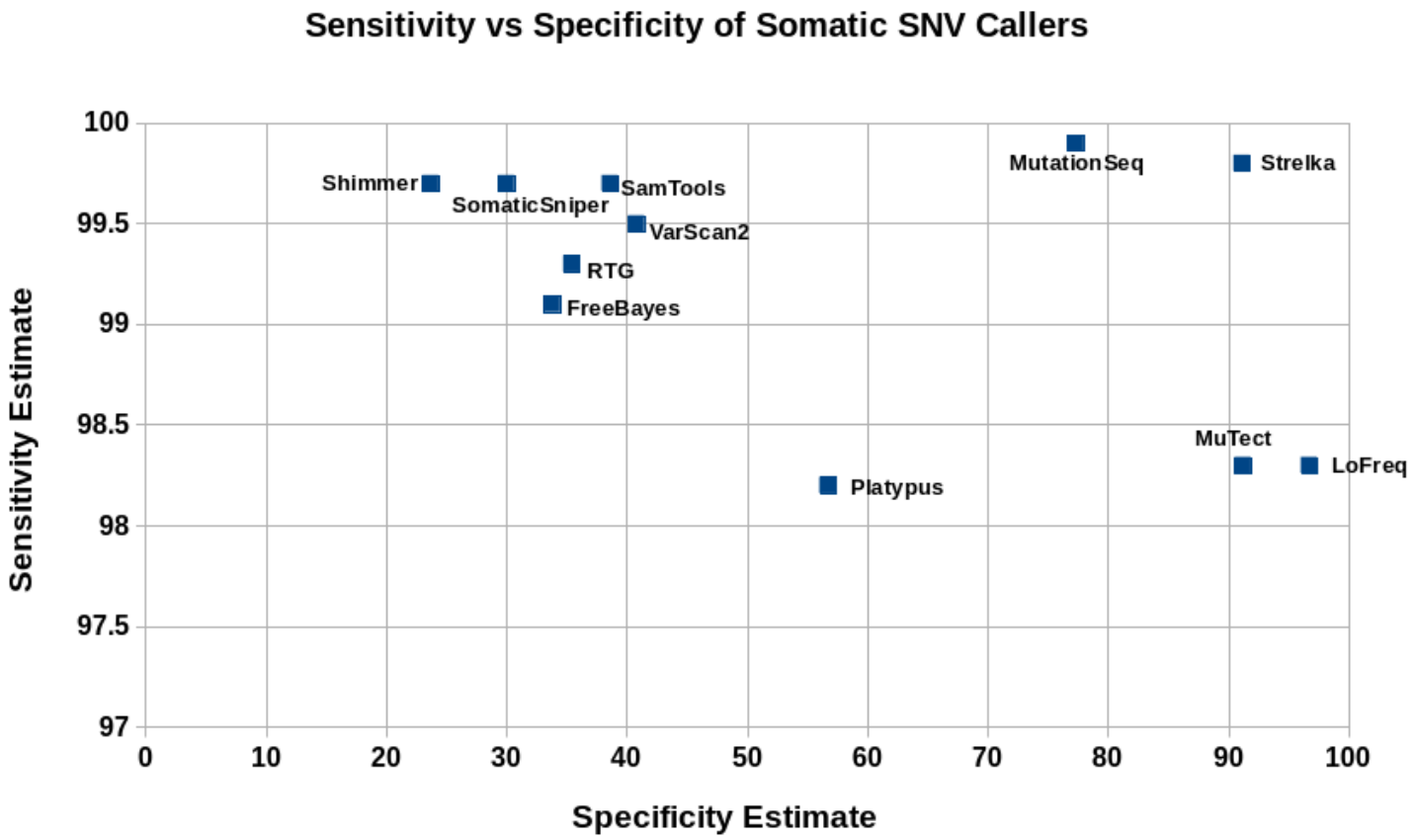
- Wall clock run times
- Number of somatic SNVs called
- Estimate sensitivity/specificity

Wall Clock Run Times



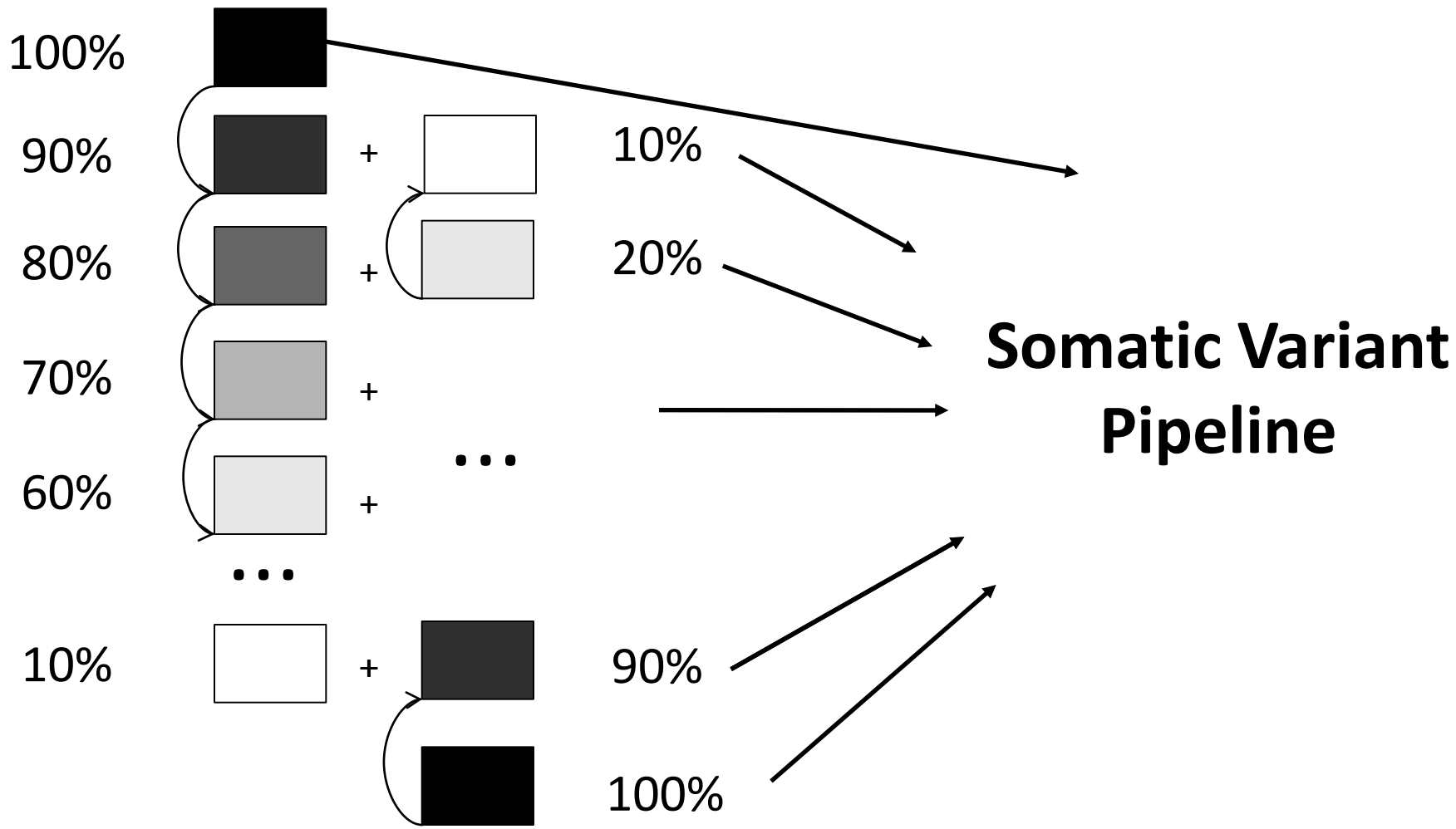
* VarScan2 requires pileup files as input, pileup generation time not included
* Single job was submitted for Strelka, but Strelka algorithm automatically splits jobs on chromosomes

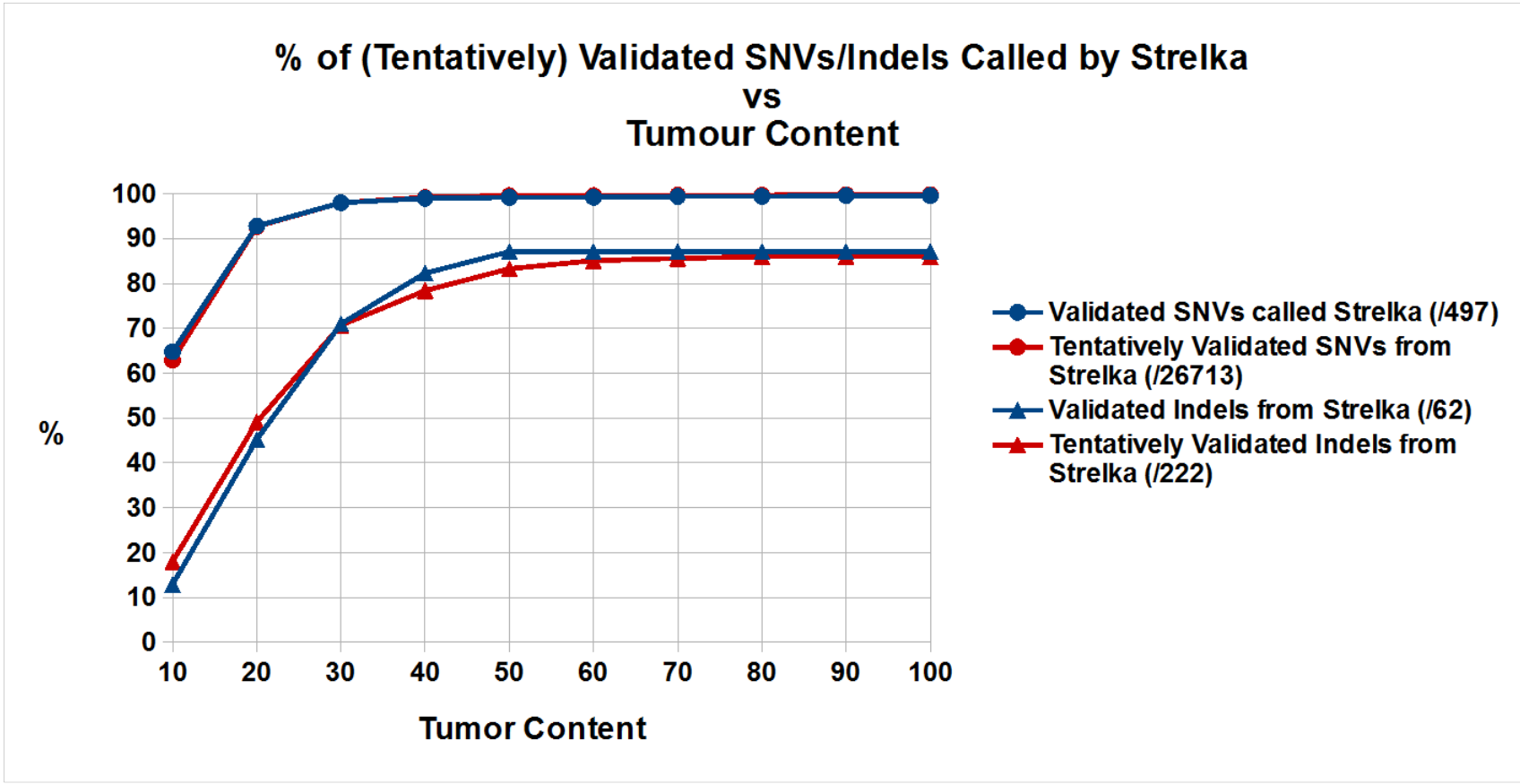

# of Somatic SNVs Called
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
LoFreq
MuTect
Shimmer
FreeBayes
Platypus
SamTools
MutationSeq
Strelka
SomaticSniper
VarScan2
RTG


Sensitivity vs Specificity of Somatic SNV Callers
Sensitivity Estimate
Specificity Estimate
100
99.5
99
98.5
98
97.5
97
0
10
20
30
40
50
60
70
80
90
100
Shimmer
SomaticSniper
SamTools
VarScan2
RTG
FreeBayes
MutationSeq
Strelka
MuTect
LoFreq
Platypus

- How does a biopsy's tumour content impact the sensitivity/specificity of predicting somatic SNVs/indels?
  - Low tumour content → miss somatic SNVs/indels

- Performed a bioinformatics titration of our in-house COLO-829 tumour/normal sample
  - COLO-829 BAM vs COLO-829BL BAM


Tumor Content
Normal Content
100%
90%
80%
70%
60%
10%
+
10%
20%
90%
100%
Somatic Variant Pipeline



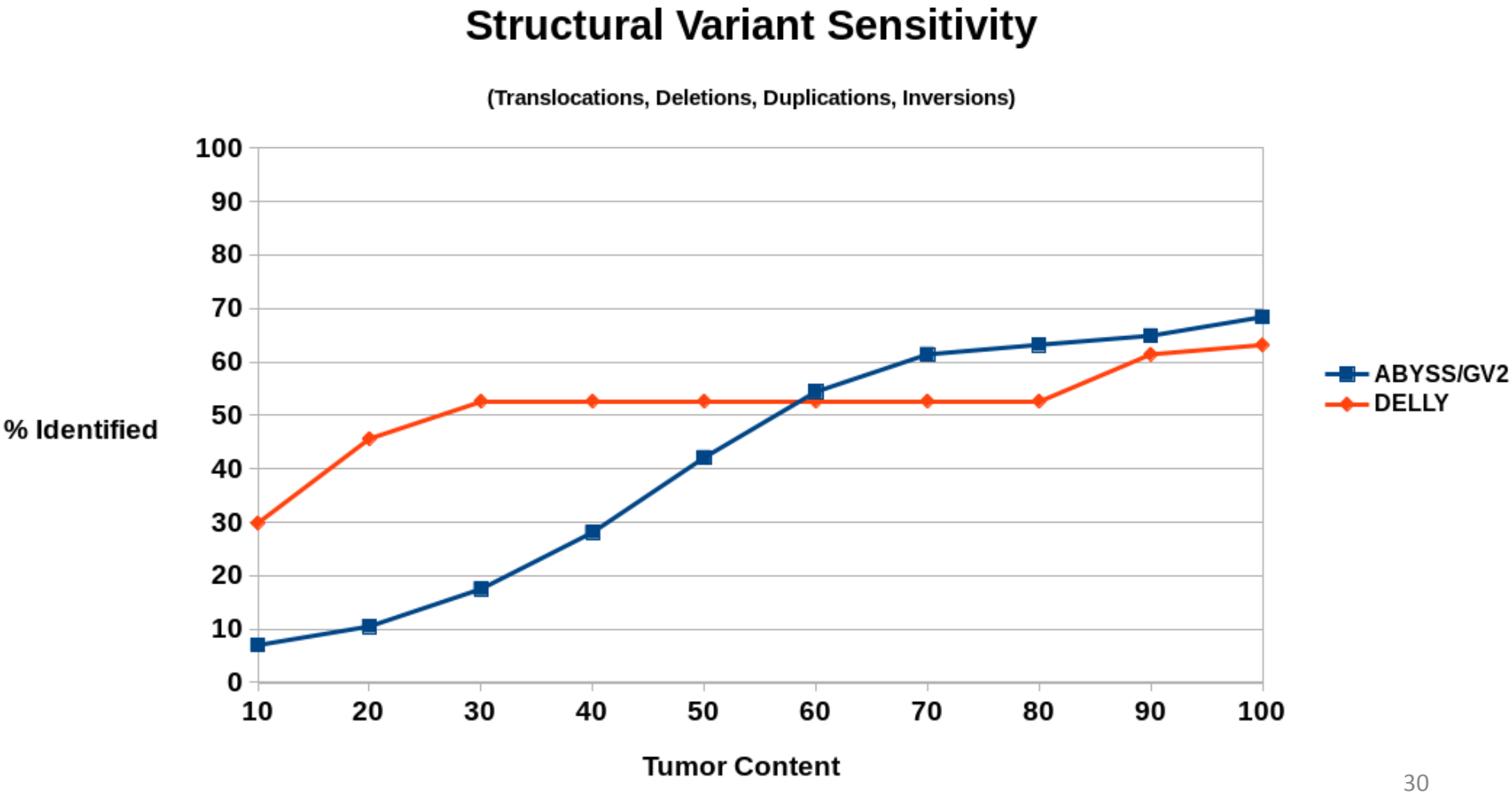
What is the minimum tumor content in which 95% of (tentatively) validated SNVs are called?

Tumor content of ~25%

Given a sample with tumor content of 40%, what percent of (tentatively) validated SNVs are called?

≥ 99%

~86% (tentatively) validated somatic indels are called at tumour content of 100%


Structural Variant Sensitivity
(Translocations, Deletions, Duplications, Inversions)
% Identified
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
70
80
90
100
Tumor Content
ABYSS/GV2
DELLY

Sci Rep. 2016 Apr 20;6:24607. doi: 10.1038/srep24607.

A somatic reference standard for cancer genome sequencing.

Craig DW[1], Nasser S[1], Corbett R[2], Chan SK[2], Murray L[3], Legendre C[1], Tembe W[1], Adkins J[1], Kim N[4], Wong S[1], Baker A[1], Enriquez D[1], Pond S[4], Pleasance E[2], Mungall AJ[2], Moore RA[2], McDaniel T[4], Ma Y[2], Jones SJ[2], Marra MA[2], Carpten JD[1], Liang WS[1].

Author information

Abstract

Large-scale multiplexed identification of somatic alterations in cancer has become feasible with next generation sequencing (NGS). However, calibration of NGS somatic analysis tools has been hampered by a lack of tumor/normal reference standards. We thus performed paired PCR-free whole genome sequencing of a matched metastatic melanoma cell line (COLO829) and normal across three lineages and across separate institutions, with independent library preparations, sequencing, and analysis. We generated mean mapped coverages of 99X for COLO829 and 103X for the paired normal across three institutions. Results were combined with previously generated data allowing for comparison to a fourth lineage on earlier NGS technology. Aggregate variant detection led to the identification of consensus variants, including key events that represent hallmark mutation types including amplified BRAF V600E, a CDK2NA small deletion, a 12 kb PTEN deletion, and a dinucleotide TERT promoter substitution. Overall, common events include >35,000 point mutations, 446 small insertion/deletions, and >6,000 genes affected by copy number changes. We present this reference to the community as an initial standard for enabling quantitative evaluation of somatic mutation pipelines across institutions.

PMID: 27094764 PMCID: PMC4837349 DOI: 10.1038/srep24607

Tool Evaluations

- Strategy to evaluate candidate tools for inclusion in production pipelines:
  - Literature search
  - Local installation of tools
  - Construct ground truth set
  - Compare results with respect to evaluation criteria
  - Choose tool(s) to use

Key Tools in Production Pipelines

- **Alignment**
  - DNA: bwa-mem
  - RNA: JAGuar, switching to STAR
  - Bisulfite: Novoalign
- **Single sample SNV / indel:** samtools mpileup
- **CNV:** cnaseq
- **LOH:** APOLLOH
- **SVs for genomes:** ABySS / DELLY / Manta
- **SVs for transcriptomes:** Trans-ABySS / DeFuse (chimerascan is being evaluated)
- **ChIP:** FindPeaks (MACS2 is being evaluated)

Conclusions

- Constructed a ground truth set of somatic SNVs/indels in COLO-829.
- Estimated sensitivity and specificity of in-house somatic SNV/indel pipeline.
- Investigated somatic SNV callers and compared their run times, number of somatic SNVs called, and accuracies.
- Even at low tumour contents (~20%), a significant number of somatic SNVs/indels were predicted accurately.

BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Acknowledgements

BCGSC:

Marco Marra

Steven Jones

Yussanne Ma

Richard Moore

Andrew Mungall

Richard Corbett

Karen Mungall

Tina Wong

Erin Pleasance

TGEN:

John Carpten

Daniel Craig

Winnie Liang

Sara Nassar

Waibhav Tembe

Christophe Legendre

Jonathan Adkins

Shukmei Wong

Angela Baker

Daniel Enriquez

Illumina:

Lisa Murray

Stephanie Pond

Nancy Kim

Timothy McDaniel

Extra Slides

- Sanger sequencing is an effective way to verify somatic mutations
  - Draw backs:
    - Laborious and $ to validate all somatic mutations identified in an NGS experiment
    - Not suitable for low frequency variants

Software	Year Published	Organization	Notes
LoFreq	2012	Genome Institute of Singapore	Bernoulli trial, assume each base is independent with sequnce error (quality score), Posisson-bionmial distribution
MuTect	2013	Broad	Bayesian classifier
Shimmer	2013	NHGRI/NIH	Fisher's exact test comparing ref/alt alleles in tumor/normal with multiple testing
FreeBayes	2012	Boston College	Bayesian statistics
Platypus	2014	The Wellcome Trust Centre for Human Genetics	Local assembly, haplotype-based, multi-sample variant caller using Bayesian statistics
SAMTools	2009	Sanger/Broad	Calculate genotype from Bayesian prior probability

Software	Year Published	Organization	Notes
SomaticSniper	2011	WUSTL	Build genotype likelihood model of MAQ, calculates probablitity of genotype differences
VarScan2	2012	WUSTL	Fisher's exact test comparing ref/alt alleles in tumor/normal
RTG Somatic	2015	Real Time Genomics, Inc	Bayesian statistics
Strelka	2012	Illumina, Inc	Bayesian statistics
MutationSeq	2011	BC Cancer Research Centre	Feature based classifiers

Sequencers at the Genome Sciences Centre

	Bases Per Second	# Machines	Total Bases / Sec.
HiSeq X	8,700,000	5	43.5 million
HiSeq 2500	3,100,000	4	12.4 million
NextSeq	1,300,000	2	2.6 million
MiSeq	50,000	3	150 thousand

~55 million bases per second

How much sequence is that?

- Human Genome : 3,000,000,000 bases (approx.)
- At the Genome Sciences Centre, we can sequence 1 human genome every:
  - 3 billion bases / 55 million bases per sec = **54.5 sec**
- The first human genome draft sequence took roughly 10 years to sequence and assemble

How do we extract meaning from the sequence data?

2,000,000,000 reads per sample

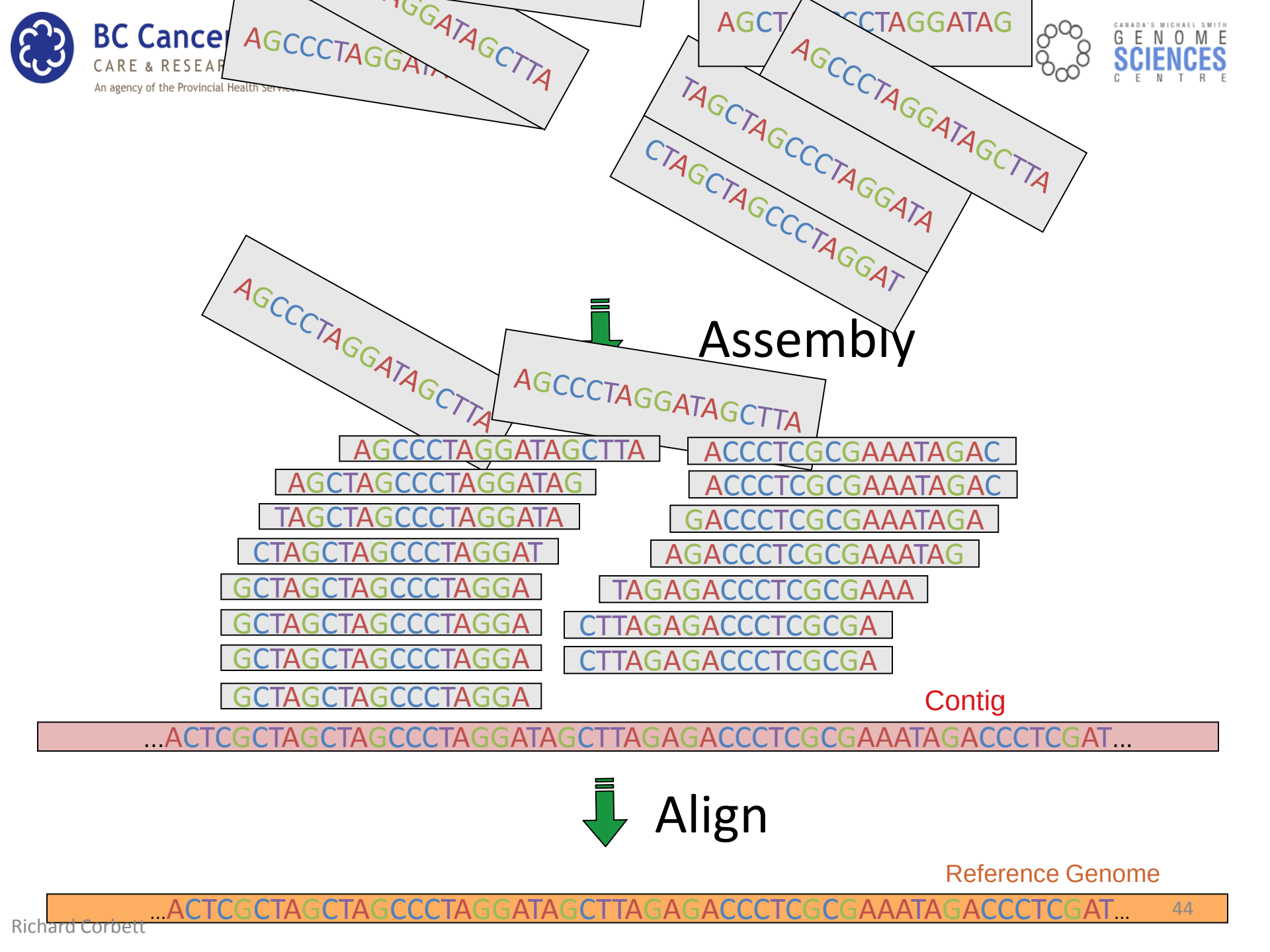
150 bases per read

3,000,000,000 base reference genome

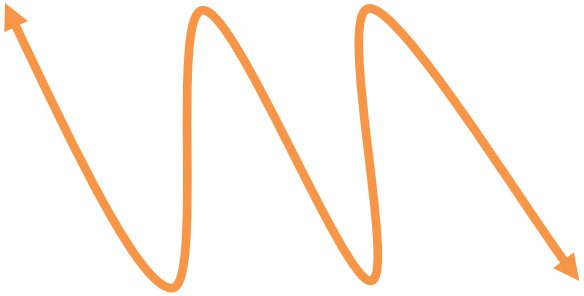
Data Interpretation

- For efficiency and to help interpretation, we often describe a sample by how it differs from a **reference sample**

- To compare samples, we
  - align sequence reads for a sample to a **reference genome**
  - find locations where our sample differs from the reference


AGCCCTAGGATAGCTTA
AGCCCTAGGATAGCTTA
AGCTAGCCCTAGGATAG
TAGCTAGCCCTAGGATA
CTAGCTAGCCCTAGGAT
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
ACCCTCGCGAAATAGAC
ACCCTCGCGAAATAGAC
GACCCTCGCGAAATAGA
AGACCCTCGCGAAATAG
TAGAGACCCTCGCGAAA
CTTAGAGACCCTCGCGA
CTTAGAGACCCTCGCGA
AGCCCTAGGATAGCTTA
AGCCCTAGGATAGCTTA
TAGCTAGCCCTAGGATA
CTAGCTAGCCCTAGGAT
AGCCCTAGGATAGCTTA
Assembly
Contig
...ACTCGCTAGCTAGCCCTAGGATAGCTTAGAGACCCTCGCGAAATAGACCCTCGAT...
Align
Reference Genome
...ACTCGCTAGCTAGCCCTAGGATAGCTTAGAGACCCTCGCGAAATAGACCCTCGAT...

Genome and Transcriptome



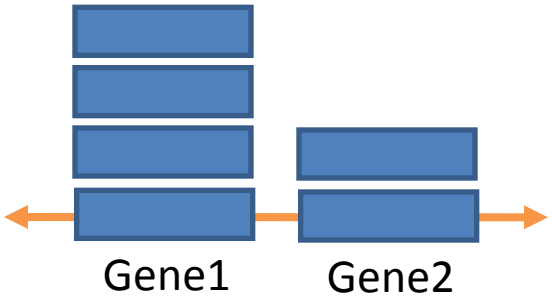
Genome sequencing allow us to find:

- SNVs (single nucleotide variants)

  CCCTTTTGGGGAA

- CNVs (copy number variants)

- SVs (structural variants)



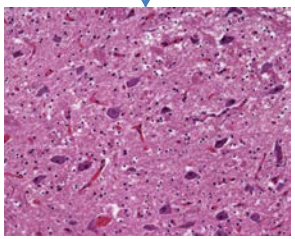
The transcriptome can be sequenced to find:

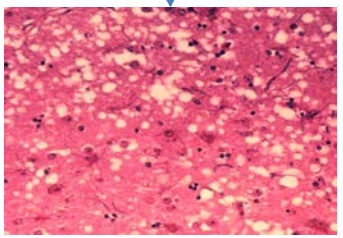
- Gene expression estimates
- Gene fusions

Personalized Medicine

Normal

Tumour (melanoma)





40X genome coverage

80X genome coverage
and
transcriptome library

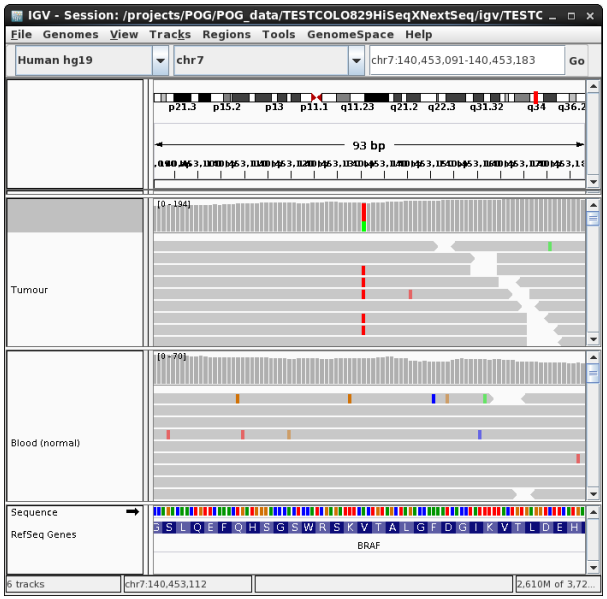
Bioinformatics

BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

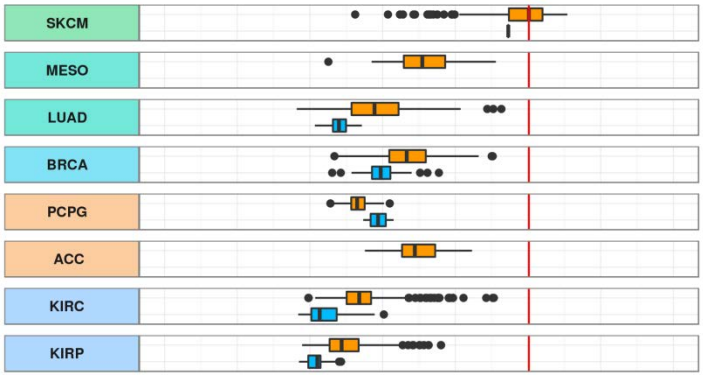
Personalized Medicine Intermediate Results

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

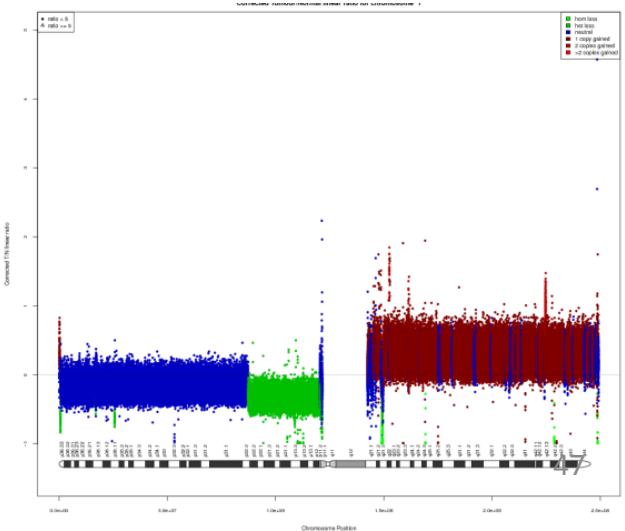
Somatic SNV calling



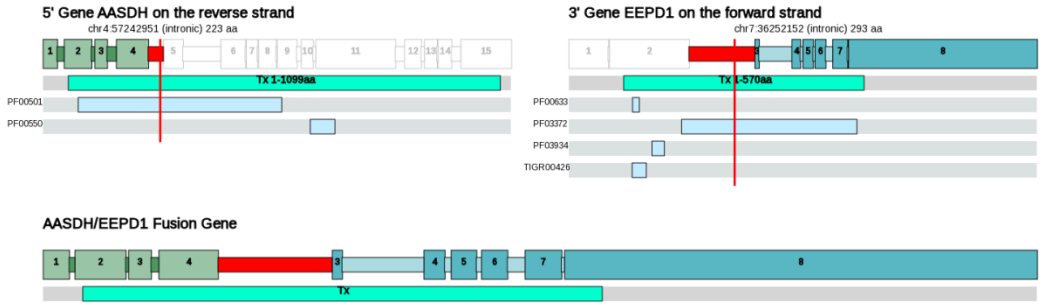
RNA expression Correlation



Somatic Copy Number



Gene fusion analysis



Richard Corbett