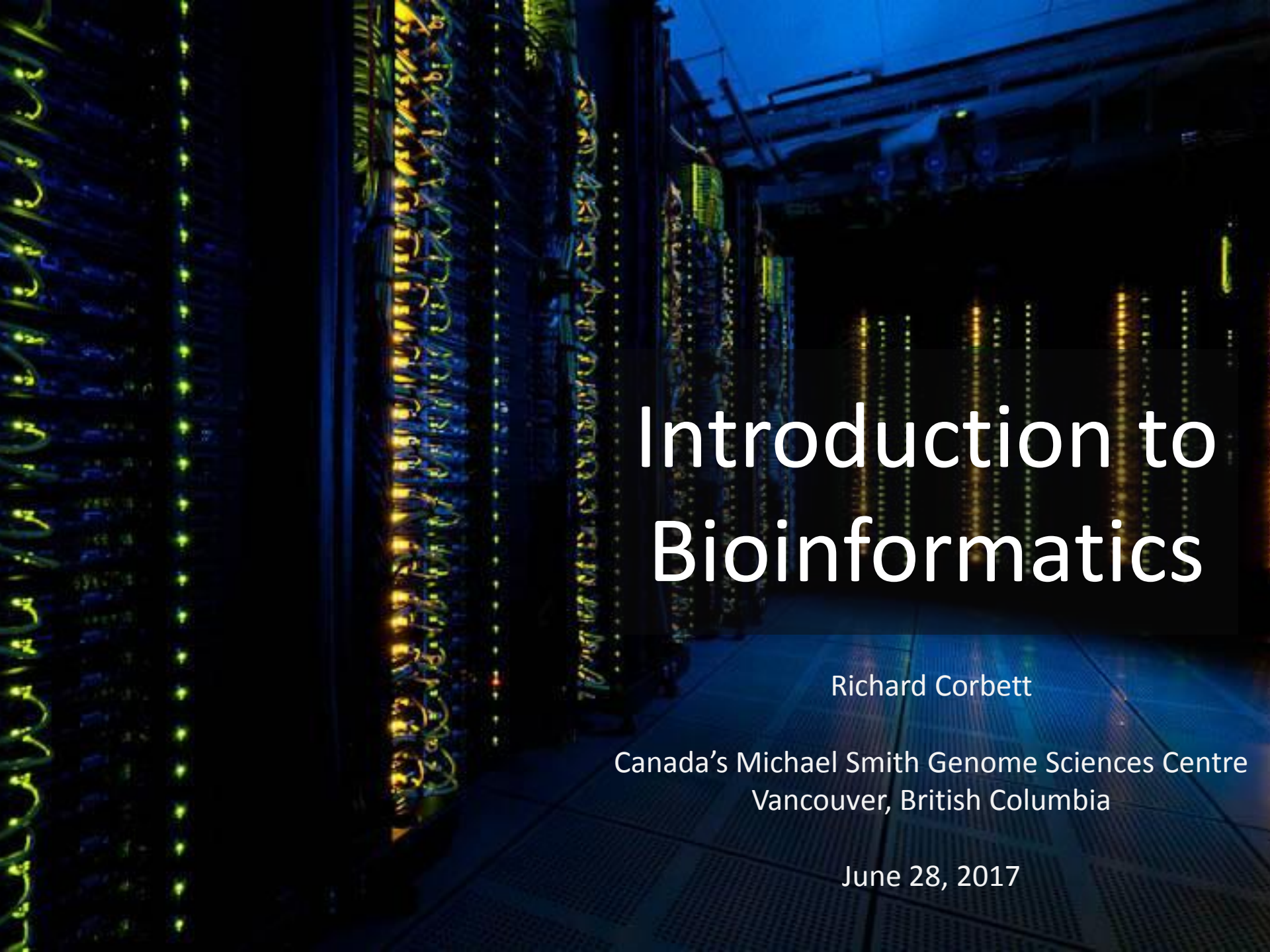
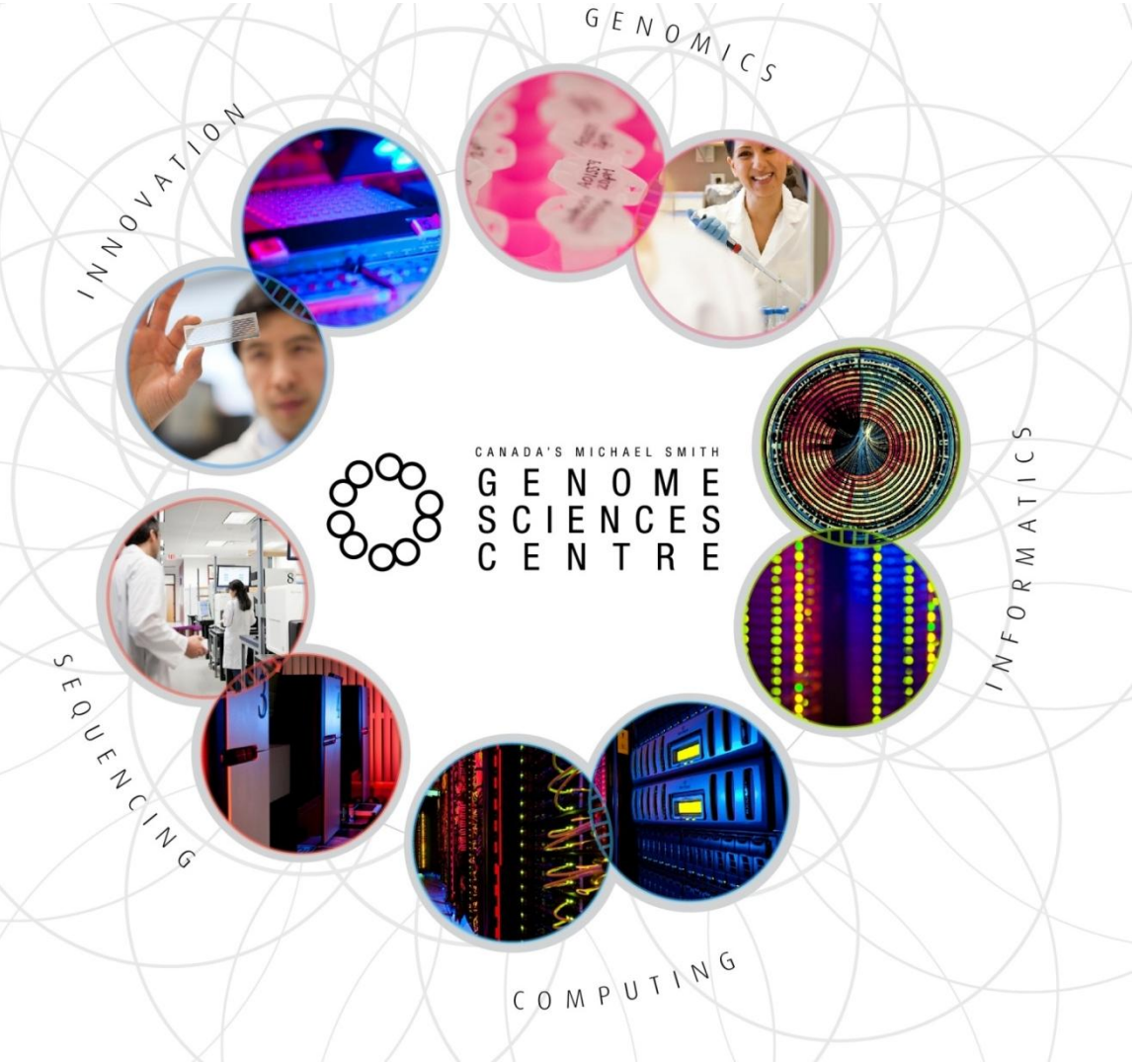


Introduction to Bioinformatics

Richard Corbett

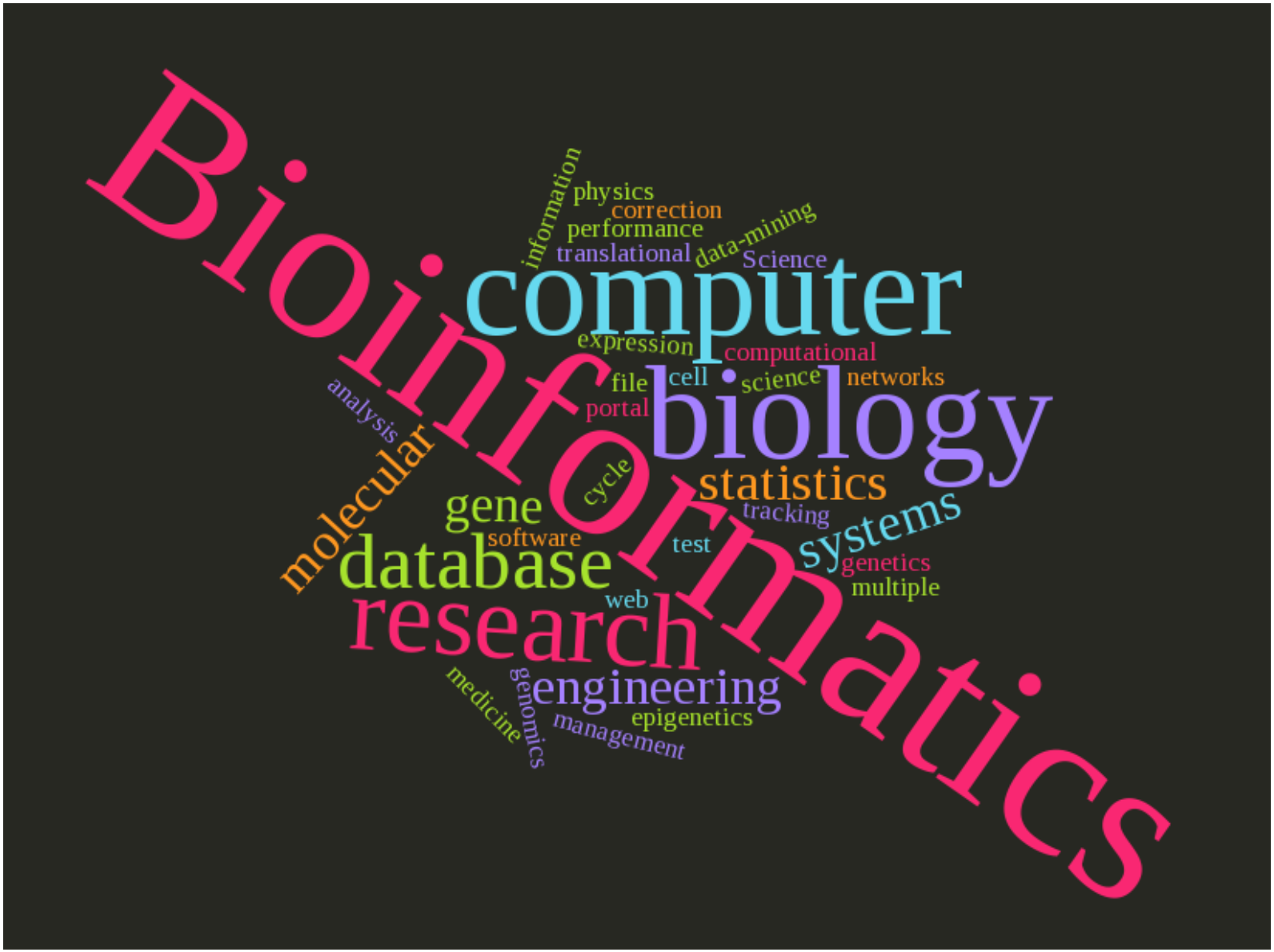
Canada's Michael Smith Genome Sciences Centre
Vancouver, British Columbia

June 28, 2017



Our mandate is to advance knowledge about cancer and other diseases and to use our technologies to improve health through disease prevention, diagnosis, and therapeutic approaches.

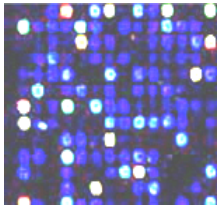
As a Process Development Coordinator I help ensure our laboratory and analytical approaches are providing the best possible results.


Bioinformatics
computer
biology
research
database
engineering
statistics
systems
gene
molecular
analysis
information
physics
correction
performance
translational
data-mining
Science
expression
computational
file
cell
science
networks
portal
cycle
tracking
software
test
genetics
multiple
web
medicine
genomics
epigenetics
management

Bioinformatics in Practice

GATCCTGGAGGATCCTGGGAG
GATCCTGGAGGATCCTGGGAG
GATCCTGGAGAATCCTGGGAG
GATCCTGGAGAATCCTGGGAG

Sequence




Other experiments

Computational models

Bioinformatics

nature

Scientific publications

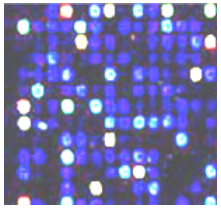
Improved disease understanding and treatment

Environmental discoveries

Bioinformatics in Practice

GATCCTGGAGGATCCTGGGAG
GATCCTGGAGGATCCTGGGAG
GATCCTGGAGAATCCTGGGAG
GATCCTGGAGAATCCTGGGAG

Sequence



Other experiments

Computational models

Bioinformatics

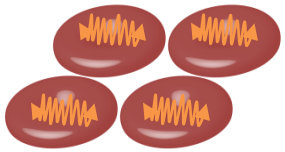
nature

Scientific publications

Improved disease understanding and treatment

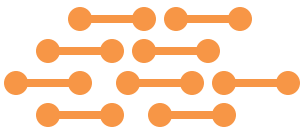
Environmental discoveries

Illumina Sequencing



1.) Cells

2.) DNA



3.) Sheared DNA, with sequencing adapters

4.) Sequencing

5.) Ready for bioinformatics

AAAAAAAAAAAAAAAACCCAAA
CCCTTTTTGGGGAAGGGGGGGTT
TCCCCCCCCCCCCCCCAAAAAAAT
AAAGGGAAAGGGGTTTCCCAAA

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 the number of bases in 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

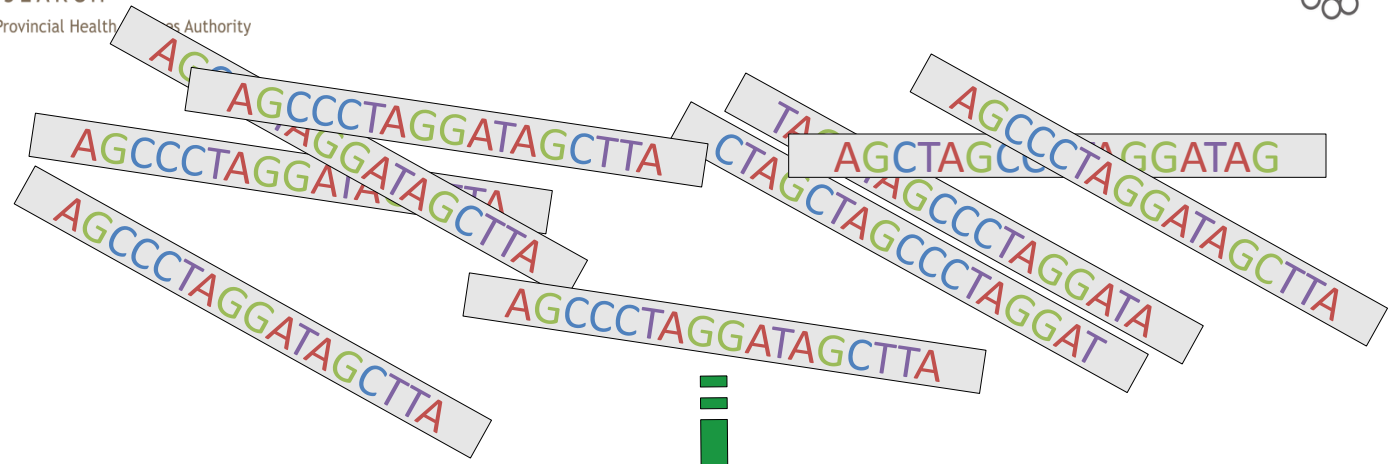
150 bases per read

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 to a **reference genome**
  - find locations where our sample differs from the reference



Alignment

Reference Genome

...ACTCGCTAGCTAGCCCTAGGATAGCTTAGAGACCCTCGCGAAATAGACCCTCGAT...

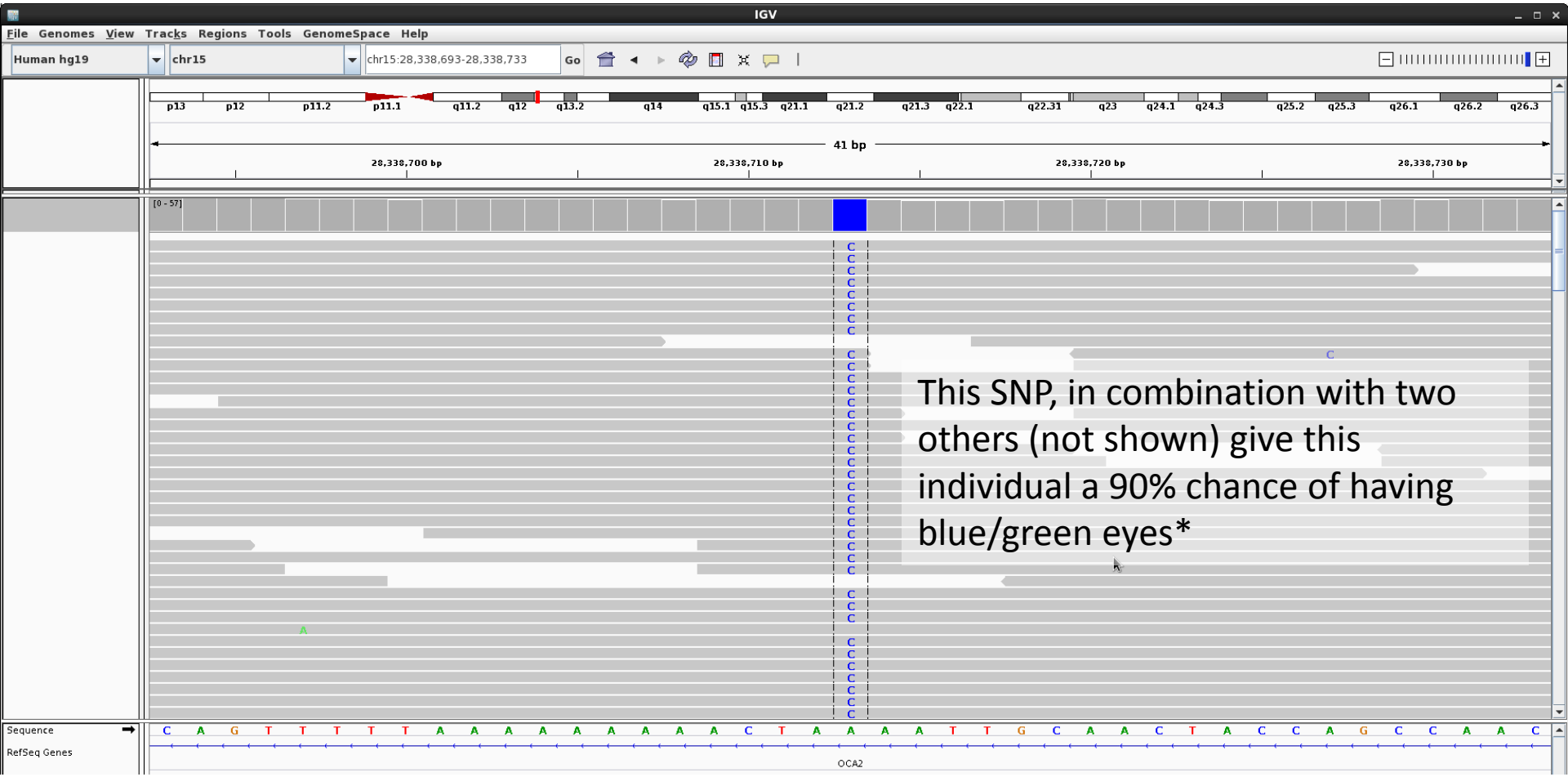
AGCCCTAGGATAGCTTA
AGCTAGCCCTAGGATAG
TAGCTAGCCCTAGGATA
CTAGCTAGCCCTAGGAT
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA
GCTAGCTAGCCCTAGGA

ACCCTCGCGAAATAGAC
ACCCTCGCGAAATAGAC
GACCCTCGCGAAATAGA
AGACCCTCGCGAAATAG
TAGAGACCCTCGCGAAA
CTTAGAGACCCTCGCGA
CTTAGAGACCCTCGCGA

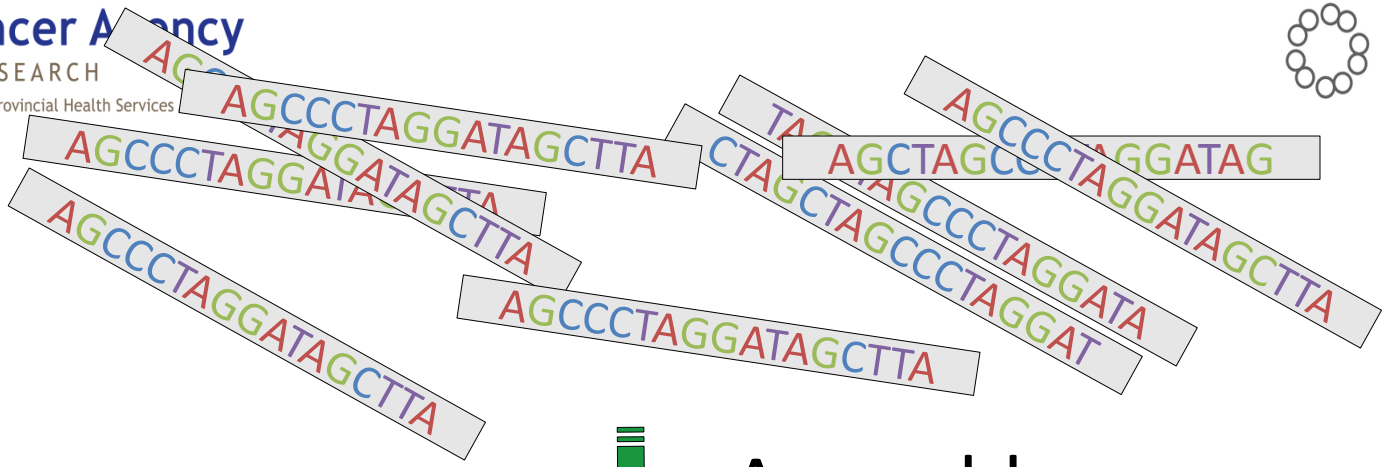
BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

Real Alignments

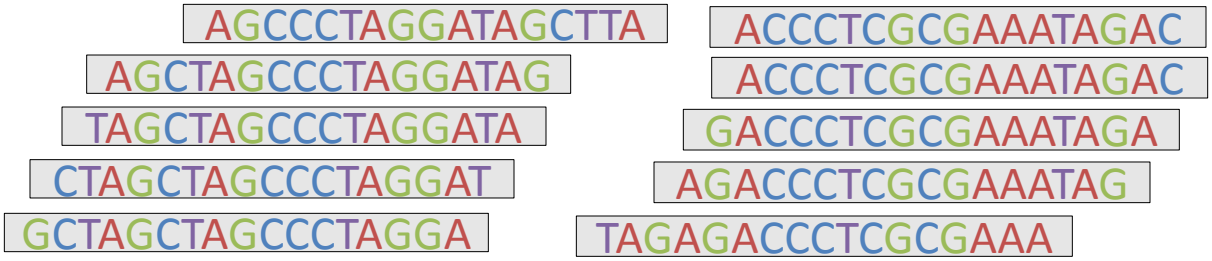
CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE



*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.

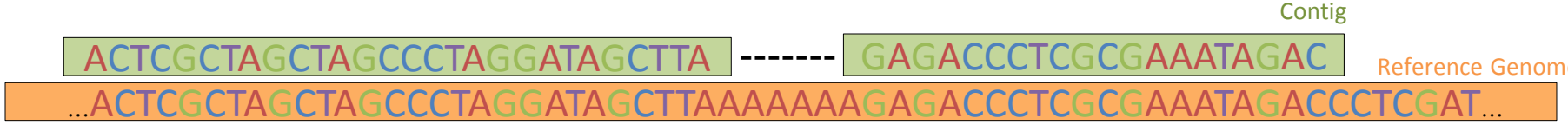


Assembly



...GCTAGCTAGCCCTAGGATAGCTTAGAGACCCTCGCGAAATAGAC... Contig

Alignment



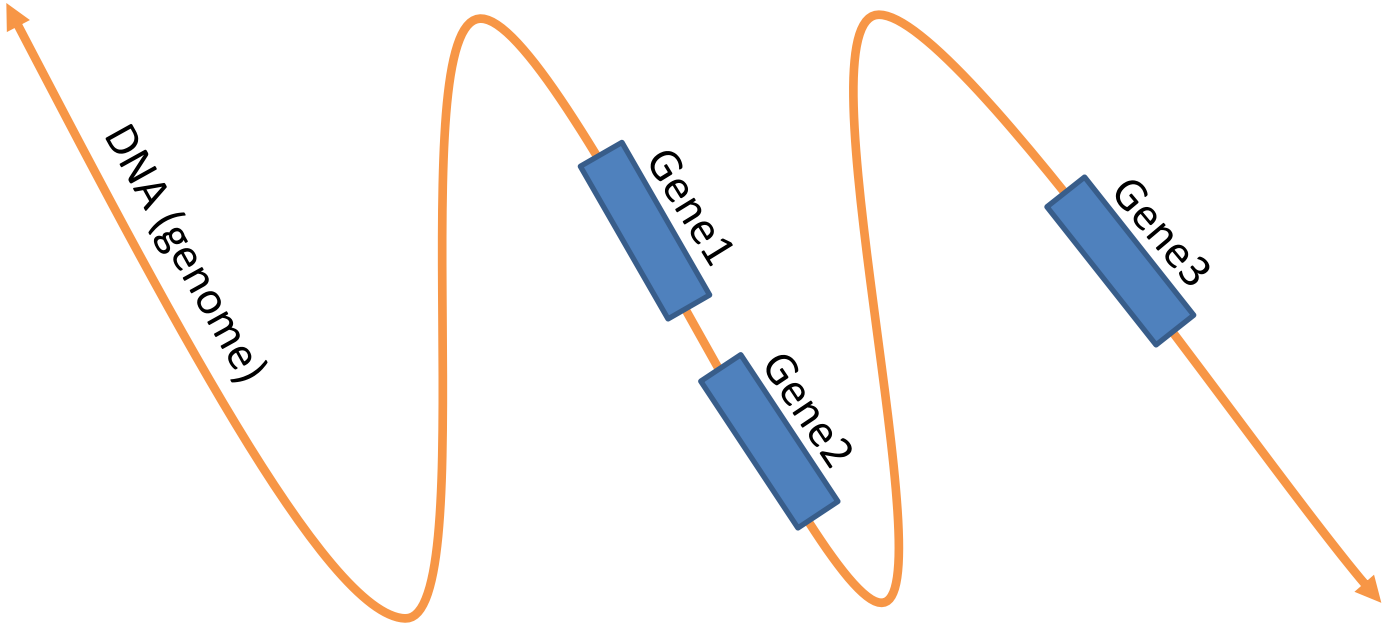
Summary So Far

- We sequence billions of reads per genome sample
- Useful / actionable results are identified via:
  - Read alignment
  - Read assembly
- We describe samples by how they differ from a reference sample

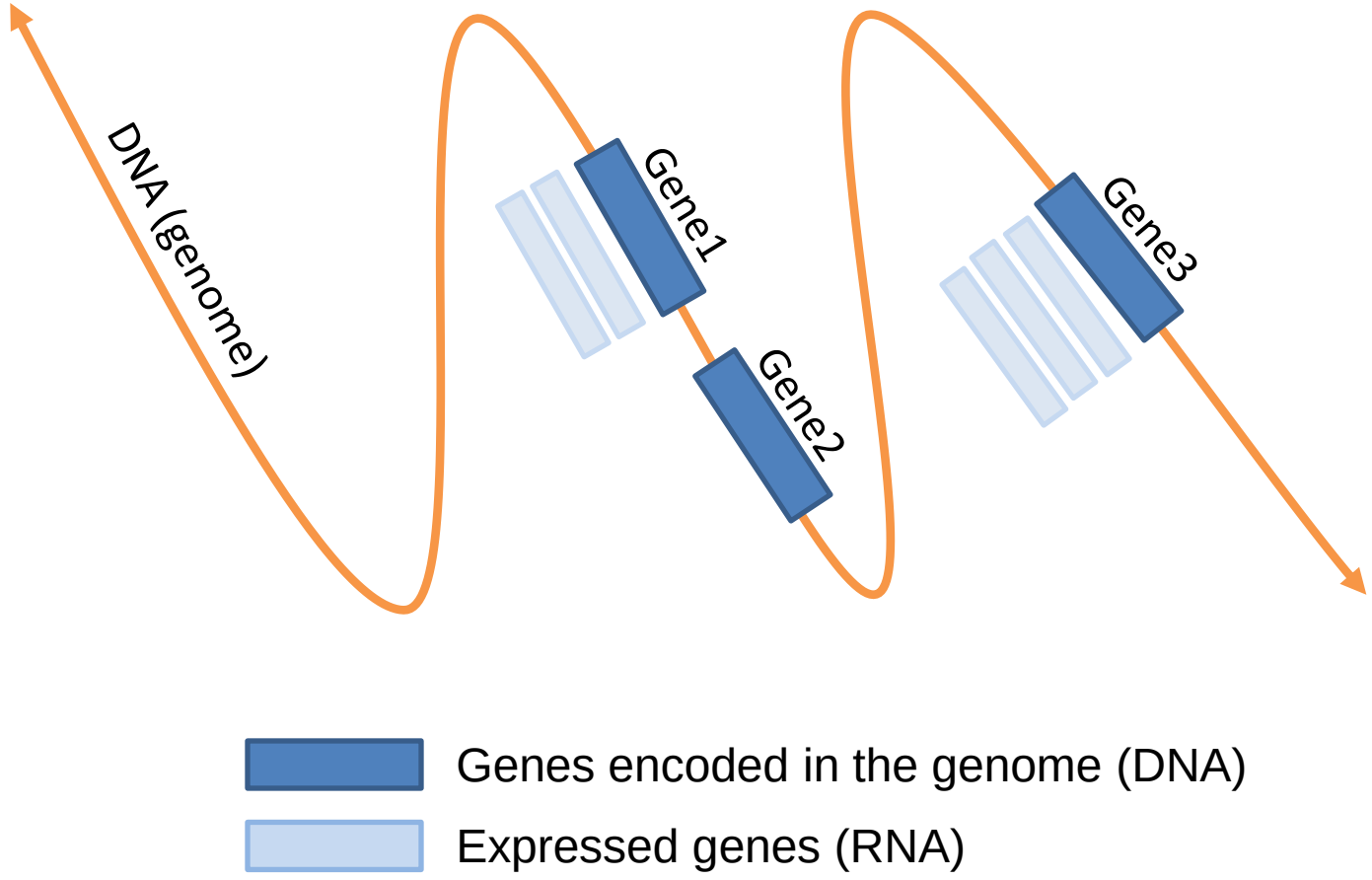
Other Data Types

- **Epigenomics** – investigates the chromatin and methylation status of regions of the genome
- **Proteomics** – measures protein expression and modifications
- **Exome/Capture** – queries specific targeted regions of the genome
- **RNA sequencing** – measures RNA expression and variation

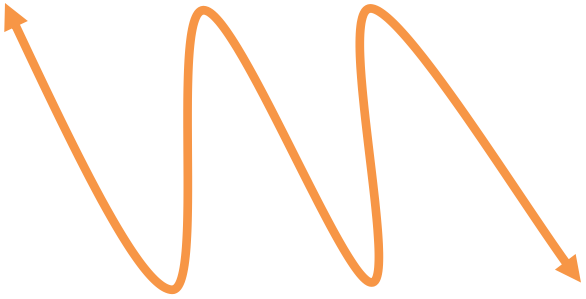
Genome and Transcriptome



Genome and Transcriptome



Genome and Transcriptome



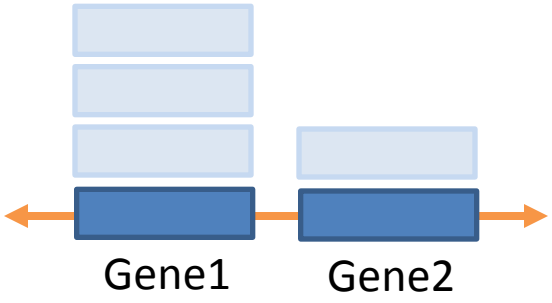
Genome sequencing allow us to find:

- SNVs (single nucleotide variants)

  CCCTTTTGGGGAA

- CNVs (copy number variants)

- SVs (structural variants)



The transcriptome can be sequenced to find:

- Gene expression estimates
- Gene fusions 

- SNVs in expressed genes

Personalized Medicine

Normal

Tumour (melanoma)

40X genome coverage

80X **genome** coverage
and
transcriptome library

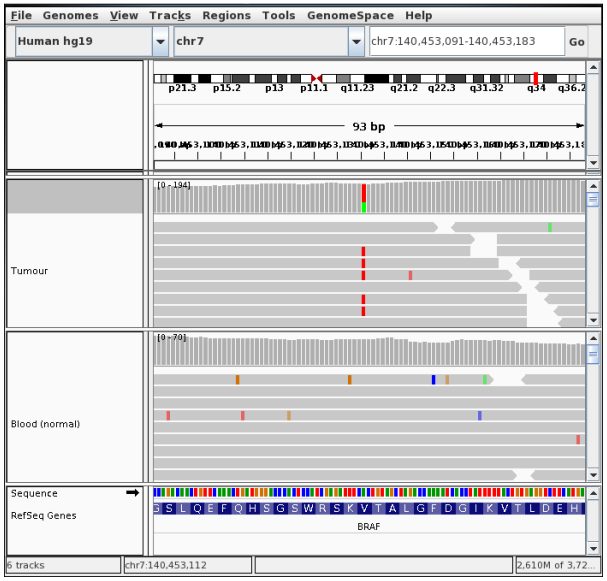
Bioinformatics

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

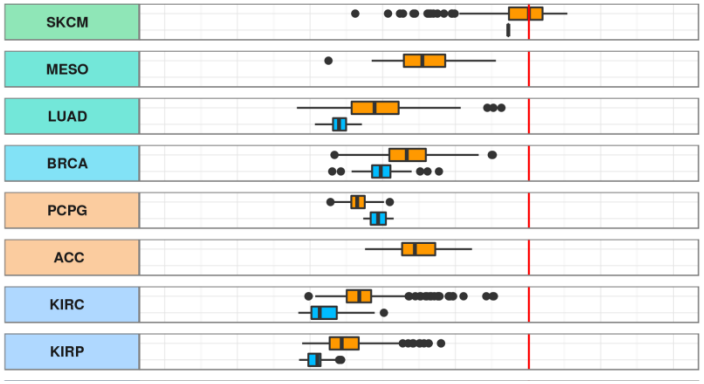
Personalized Medicine Intermediate Results

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

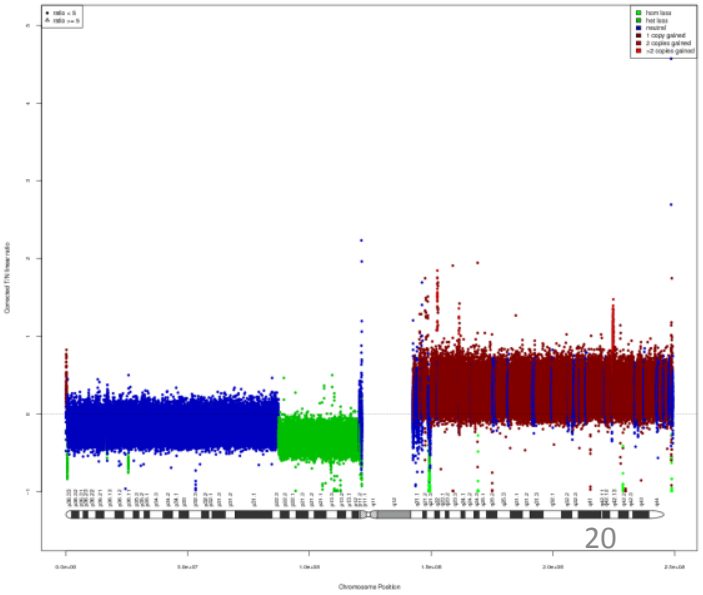
Somatic SNV calling



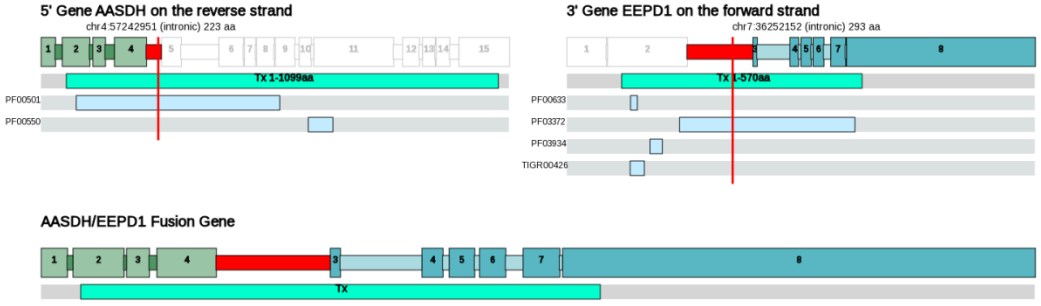
RNA expression correlation



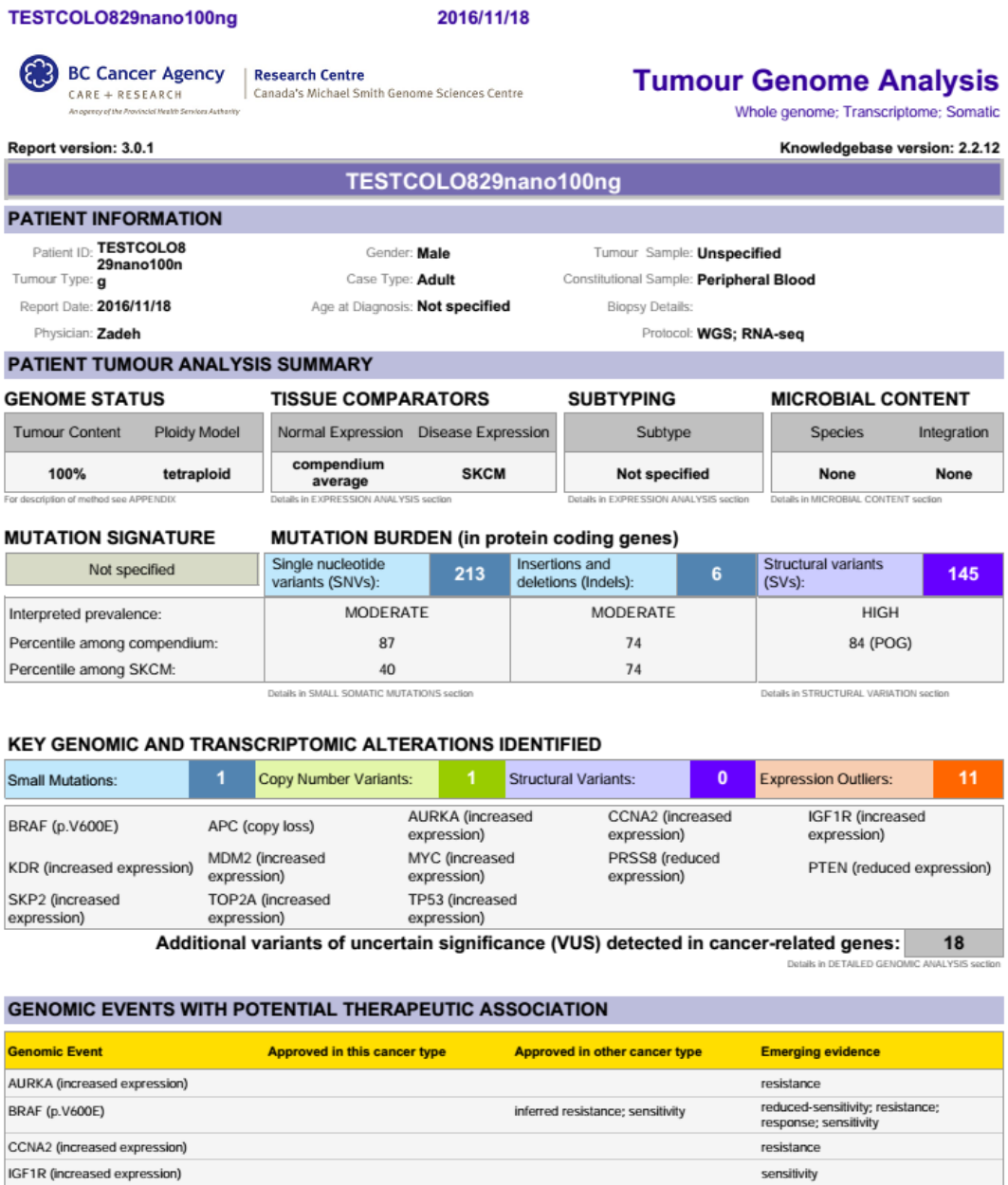
Somatic copy number



Gene fusion analysis



Report



TESTCOLO829nano100ng 2016/11/18

BC Cancer Agency | Research Centre
CARE + RESEARCH | Canada's Michael Smith Genome Sciences Centre
An agency of the Provincial Health Services Authority

Tumour Genome Analysis
Whole genome; Transcriptome; Somatic

Report version: 3.0.1 Knowledgebase version: 2.2.12

TESTCOLO829nano100ng

PATIENT INFORMATION

Patient ID:	TESTCOLO829nano100ng	Gender:	Male	Tumour Sample:	Unspecified
Tumour Type:		Case Type:	Adult	Constitutional Sample:	Peripheral Blood
Report Date:	2016/11/18	Age at Diagnosis:	Not specified	Biopsy Details:	
Physician:	Zadeh			Protocol:	WGS; RNA-seq

PATIENT TUMOUR ANALYSIS SUMMARY

GENOME STATUS		TISSUE COMPARATORS		SUBTYPING	MICROBIAL CONTENT	
Tumour Content	Ploidy Model	Normal Expression	Disease Expression	Subtype	Species	Integration
100%	tetraploid	compendium average	SKCM	Not specified	None	None
For description of method see APPENDIX		Details in EXPRESSION ANALYSIS section		Details in EXPRESSION ANALYSIS section	Details in MICROBIAL CONTENT section	

MUTATION SIGNATURE

Not specified

MUTATION BURDEN (in protein coding genes)

	Single nucleotide variants (SNVs): 213	Insertions and deletions (Indels): 6	Structural variants (SVs): 145
Interpreted prevalence:	MODERATE	MODERATE	HIGH
Percentile among compendium:	87	74	84 (POG)
Percentile among SKCM:	40	74	

Details in SMALL SOMATIC MUTATIONS section | Details in STRUCTURAL VARIATION section

KEY GENOMIC AND TRANSCRIPTOMIC ALTERATIONS IDENTIFIED

Small Mutations: 1	Copy Number Variants: 1	Structural Variants: 0	Expression Outliers: 11

BRAF (p.V600E)	APC (copy loss)	AURKA (increased expression)	CCNA2 (increased expression)	IGF1R (increased expression)
KDR (increased expression)	MDM2 (increased expression)	MYC (increased expression)	PRSS8 (reduced expression)	PTEN (reduced expression)
SKP2 (increased expression)	TOP2A (increased expression)	TP53 (increased expression)		

Additional variants of uncertain significance (VUS) detected in cancer-related genes: 18

Details in DETAILED GENOMIC ANALYSIS section

GENOMIC EVENTS WITH POTENTIAL THERAPEUTIC ASSOCIATION

Genomic Event	Approved in this cancer type	Approved in other cancer type	Emerging evidence
AURKA (increased expression)			resistance
BRAF (p.V600E)		inferred resistance; sensitivity	reduced-sensitivity; resistance; response; sensitivity
CCNA2 (increased expression)			resistance
IGF1R (increased expression)			sensitivity

A detailed report of results for a patient's sample is provided to the clinician allowing them to view the genetic landscape of a patient's disease.

This approach has enabled treating clinicians to make informed clinical decisions based on the genomic information integrated with other clinical features.

Summary

- Bioinformatics is (among other things) the process through which the interpretation of billions of sequence observations yields a distilled list of actionable findings

- This is accomplished in equal parts by:
  - Powerful computing infrastructure
  - Advanced algorithms
  - Trained individuals from a diverse set of fields

Upcoming webinars & training events can be found on the WestGrid website:
https://www.westgrid.ca/events

If attendees would like to learn more about Compute Canada / WestGrid high performance computing resources and services, please visit:
https://docs.computecanada.ca

computecanada

WESTGRID