Canadian Bioinformatics Workshops

www.bioinformatics.ca

@Bioinfodotca

@bioinfojobs

Facebook.com/bioinformaticsdotca

Bioinformatics-ca.github.io
Learning Objectives of Module 5

1. Be introduced to PCAWG as an use case utilizing multiple cloud resources for large-scale cancer genome analyses

2. Be aware of lessons learnt from PCAWG, and issues to consider when planning your own genomic analyses on the cloud

3. Be familiar with PCAWG resources

4. Be able to run one of the PCAWG workflows
Part 1. What is PCAWG?

Pan-Cancer Analysis of Whole Genomes (PCAWG)

• An international collaboration to identify common patterns of mutations in more than 2800 cancer whole genomes from the International Cancer Genome Consortium (ICGC)

• Goals: Understand the nature and consequences of somatic and germline variations in both coding and non-coding regions with specific emphases on
  • Non-coding RNAs & regulatory elements
  • Genomic structural alterations
  • Pathogen (viral) insertions
  • Mutation signatures
  • Tumor-specific driver pathways
Organization of PCAWG

- Steering committee
  - Peter Campbell, Wellcome Trust Sanger Institute
  - Gaddy Gatz, Broad Institute
  - Jan Korbel, European Molecular Biology Laboratory
  - Lincoln Stein, Ontario Institute for Cancer Research
  - Josh Stuart, University of California Santa Cruz

- 16 Research Working Groups: each focusing on a variant type or research question, 700 researchers and 130 abstracts

- Technical Working Group
  - responsible for uniform alignment and variant calling of >2800 pairs of whole genomes
  - data curation, quality checking and data dissemination

16 Research Working Groups

1. Novel somatic mutation calling methods
2. Analysis of mutations in regulatory regions
3. Integration of the transcriptome and genome
4. Integration of the epigenome and genome
5. Consequences of somatic mutations on pathway and network activity
6. Patterns of structural variations, signatures, genomic correlations, retrotransposons and mobile elements
7. Mutation signatures and processes
8. Germline cancer genome
9. Inferring driver mutations and identifying cancer genes and pathways
10. Translating cancer genomes to the clinic
11. Evolution and heterogeneity
12. Portals, visualization and software infrastructure
13. Molecular subtypes and classification
14. Analysis of mutations in non-coding RNA
15. Mitochondrial
16. Pathogens
Initial Roadmap for Technical WG (2014)

Data Train 1: Collect normal and tumor whole genomes from at least 2,000 donors
- unaligned lane BAMs - 150GB per WG
- amounts to 600TB of unaligned BAMs

Data owners prepare reads in the format of unaligned lane level BAMs and metadata using PCAP tool, and upload to GNOS

Align reads with BWA-Mem
- results in 600TB aligned BAMs
- Alignment started in August 2014

Variant calling using 3 core pipelines
- Sanger
- DKFZ/EMBL
- Broad/MuSE
(Initial estimates of runtime and compute requirements were very vague)

63 donors with sufficient DNA material were selected for validation
- All algorithms were invited to submit variant calls on these donors which will then be validated to determine accuracies of the algorithms
- SNVs - 10 callers
- indels - 8 callers
- SVs - 9 callers

Calls were strategically selected for deep-targeted sequencing (~9 months)

Develop consensus strategy based on validation results to merge variant calls from the 3 core pipelines

Consensus variant calls

Challenges During Alignment Phase

- Disk space - 600TB of raw data + 600TB aligned BAM
- Bandwidth
  - Host data centres in multiple regions for local uploads, ie. Europe, North America, Asia
  - Limit the movement of raw data. Use compute local to the data centre for alignment
- Reproducibility - Do alignments of the same WG performed at 2 centres yield the same results? Yes
- How to track this much data?
  - metadata standardization
  - Elasticsearch indexing
Outcome of Alignment

• ~2000 core-hours per donor

• Started running BWA-Mem in August 2014, and finished ~2000 donors by December 2014

• Data Train 2: additional 800 donors were accepted in April 2015. Alignment took place concurrently with variant calling

• Next, 3 variant calling pipelines - it should get easier, right?
Challenges During Variant Calling Phase

• Pipelines were well tuned to author’s own compute environment, eg. SGE, Firehose, but porting them to run on cloud requires some code development

• Pipelines were well tuned to specific sequencing centres, but were naive to library artifacts, read group naming schemes from other centres

• Some pipelines have newly developed algorithms

• More compute is required

Components in the 3 Core Variant Calling Pipelines

<table>
<thead>
<tr>
<th></th>
<th>Sanger</th>
<th>DKFZ/EMBL</th>
<th>Broad/MuSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download BAMs</td>
<td>gtdownload</td>
<td>gtdownload</td>
<td>gtdownload</td>
</tr>
<tr>
<td>SNVs</td>
<td>CaVEMan</td>
<td>dkfz_snv</td>
<td>MuTect, MuSE</td>
</tr>
<tr>
<td>Indels</td>
<td>Pindel</td>
<td>platypus</td>
<td>Snowman</td>
</tr>
<tr>
<td>SVs</td>
<td>Brass</td>
<td>DELLY</td>
<td>dRanger, Snowman</td>
</tr>
<tr>
<td>CNVs</td>
<td>ASCAT</td>
<td>ACEseq</td>
<td>N/A</td>
</tr>
<tr>
<td>Germline</td>
<td>N/A</td>
<td>above components</td>
<td>HaplotypeCaller</td>
</tr>
<tr>
<td>Upload results</td>
<td>gtupload</td>
<td>gtupload</td>
<td>gtupload</td>
</tr>
</tbody>
</table>

Core-hours per donor***

<table>
<thead>
<tr>
<th></th>
<th>Sanger</th>
<th>DKFZ/EMBL</th>
<th>Broad/MuSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>800</td>
<td>800</td>
<td>2300</td>
</tr>
</tbody>
</table>

*** These are averages from actual runs on AWS. Runs are typically slower in academic environments (older CPUs, IO bottleneck, network traffic?).

A challenge at the time: run times and compute requirements were not completely clear at the start of running of these workflows.
14 More Compute Resources

- OICR, Toronto*
- AWS, Ireland
- Sanger, Hinxton
- EBI, London
- DKFZ, Heidelberg
- ETRI, Seoul
- UTokyo
- NCI Cluster, UCSC
- IDASH, UCSD
- PDC, UChicago
- Microsoft Azure
- AWS, Virginia*
- Seven Bridges Genomics
- BSC, Barcelona

* OICR’s Cancer Genome Collaboratory and Amazon Web Services started hosting PCAWG data in Nov 2015

Module 5

Core Analyses Completed

- 2834 donors, 5795 specimens
After Core Analyses, Quality Checking by Multiple Groups

176 donors are excluded (6% of 2834)
- lack of clinical or histological information
- tumor with ≥4% contamination
- normal with ≥15% tumor contamination
- excessive number of mutations as synonymous or in dbSNP
- cDNA or mouse contamination
- SV artifact
- extreme outliers based on QC metrics in 5-star rating

76 donors are greylisted (2.6% of 2834)
- low-level of contamination, filtering has rescued the samples but use with caution

2583 donors with 2778 tumors for downstream analyses
- 1188 with tumor RNA-Seq + 150 normal RNA-Seq

Artifact Filtering & Annotations

Filtered out
- Oxidative artifacts
- PCR template bias
- Forward/reverse strand bias
- Non-robust mapping
- Panel of normals
- SNVs that overlap with germline calls
- 1000 genome SNPs
- Chromosome Y calls in female donors

Annotations
- R1/R2/N3 signature artifacts - Sanger towers / T>A bleed through / C > A oxo-guanine
- enriched SNVs near indels
- tumour in normal estimates
- QC 5-star ratings
Consensus Strategy for SNVs

- 2+ of 4 callers (Sanger’s CaVEMan, DKFZ, Broad’s MuTect, WashU’s MuSE)
- ~90% precision, ~90% sensitivity

Consensus Strategy for Indels

- Stacked logistic regression: each core caller's "vote" for the call is weighted by how it has performed on calls with similar features
- 3 Core callers: Sanger’s Pindel, DKFZ’s Platypus, BSC’s SMuFin. Broad’s Snowman was used as a feature for calls present in these call sets
- 60% sensitivity 90% precision
Part 2. Lessons Learnt from PCAWG

How PCAWG Managed Multiple Clouds

1. Centralized Metadata
2. Project manager
3. Cloud shepherd
4. Cloud Orchestrator
We Need Both Academic and Commercial Clouds

- Run everything on academic clouds?
  - Low up-front cost
  - Each environment is slightly different and requires at least 0.5 FTE as cloud shepherd
  - One workflow requires 32 cores/256GB RAM. Few academic clouds have large number of such VMs

- Running everything on commercial clouds?
  - Large number of available VMs for 0.5 FTE to supervise
  - Some jurisdictions don’t allow their cohorts to be analyzed on commercial clouds
  - Some donors took >2 months to analyze. We’ve learnt to predict long-running donors and save them for academic clouds

Lessons Learnt from PCAWG

1. Metadata is key to tracking raw data and workflow outputs
   - Need to standardize and validate metadata of other data types (RNA-Seq, Bi-Sulfite) as we did for whole genomes

2. Be ready for outages/instabilities of data centers
   - Replicate data to multiple repositories
   - Develop workflows that are aware of multiple data sources and automatically retry in a preferred priority

3. Be ready for outages of compute resources
   - Queuing system needs to handle real-time changes of job reassignment from one cloud to another
   - Minimize cleanup after outages
Lessons Learnt from PCAWG

4. Use BAM statistics (coverage, discordant reads) to determine the right size of VM to use, and to predict long-running donors

5. Cost of compute is lower than an FTE
   - Build in more logic in queuing system to retry failed jobs on progressively larger VMs, before having a cloud shepherd to debug
   - Better monitoring systems with alerts, eg. idle VM

6. Build validators for workflow results before submission to data centers, just as we did for raw data. This should be extended to downstream analyses: gene expression, germline haplotypes, etc.

Future Cloud Orchestration: More Centralized

- DACO/dbGaP
- Cloud Orillator (queuing, monitoring)
- Compute algorithms
- Compute
- Commercial Clouds
- Sequencing Centers
- Genome data
- Protected Data
  - Cloud, Chicago
  - TCGA data
- Data
- Compute
- Results
- Commercial Clouds
  - Cancer Genome Collaboratory, OICR
  - AWS
- Cloud Orillator
  - Query metadata, apps store
  - Upload compute algorithms
Future Projects That Will Benefit from PCAWG’s Lessons

- Pan-Prostate Initiative
  - >1000 whole genomes, >500 exomes and more data types than PCAWG
  - Data from US, Canada, UK, Germany, Australia
  - Compute in Canada, UK and Australia

- ICGCmed
  - Link genomic data to clinical information
  - Goal: sequence samples from >200,000 cancer patients over 10 years
  - Almost 100x PCAWG

Part 3. PCAWG Resources
PCAWG Resources Available to Researchers

1. Data, data, data
   - What data is available?
   - Where is it?
   - How do you download them to the cloud?

2. Analytical methods
   - Workflows developed by Technical Working Group available as dockers
   - Methods developed by Research Working Groups will be available but may not be dockerized

1. Best practices
   - What criteria to use to exclude samples?
   - How to identify artifacts and possibly remove them?

Part 3.1. PCAWG Data
Search for Aligned BAMs & Variants at ICGC Data Portal

File Entity Page: Metadata & Other Details

Faceted search on donor or file characteristics

Unique identifier for a file that is hosted at multiple repositories
Real Time BAM Stats at Using iobio

Real time VCF Stats (coming soon)
File Entity Page: Metadata & Other Details

Unique identifier for a file that is hosted at multiple repositories

PCAWG Metadata in JSON format
Module 5: Bioinformatics.ca

Search for Aligned BAMs & Variants at ICGC Data Portal
Non-redundant manifests based on clouds order of preference

Generate manifest(s)

A manifest is a text file passed to download clients in order to download selected files of interest. Each repository's manifest file can be individually downloaded.

Duplicates files from repo lower in the list are removed

Rows can be dragged to change repo priority

An archive that contains a manifest file for each selected repository is generated

<table>
<thead>
<tr>
<th>Repository</th>
<th># Errors</th>
<th># Files</th>
<th>Total file size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGL - Toronto</td>
<td>1,569</td>
<td>45,610</td>
<td>467,321 MB</td>
</tr>
<tr>
<td>ESI - Vienna</td>
<td>1,646</td>
<td>3,413</td>
<td>407,141 MB</td>
</tr>
<tr>
<td>MPS - Virginia</td>
<td>1,402</td>
<td>33,402</td>
<td>381,597 MB</td>
</tr>
<tr>
<td>PDC - Chicago</td>
<td>885</td>
<td>3,570</td>
<td>231,317 MB</td>
</tr>
<tr>
<td>Total</td>
<td>58,002</td>
<td>90,092</td>
<td>1,46 PB</td>
</tr>
</tbody>
</table>

icgc-get: A Universal Download Client for ICGC Data

I - Search Data

II - Get a manifest-id

III - icgc-get download -m <manifest-id>

How it Works

- Uses Portal to define / serve your manifest (manifest-id)
- Checks you are authorised to access the requested data
- Downloads files from repositories in a chosen order of preference (Eliminates duplicates)
ICGC-Storage-Client
For Accessing Data from the Collaboratory cloud

Download manifest data

```
%: icgc-storage-client download --manifest 4jdyyqa099ew22
--output-dir data --output-layout bundle
```

Download BAM slices

```
%: icgc-storage-client view --object-id ea17647-17f6-5ae0
--query 12:25357723-25403870
```

Mounting a manifest (FUSE)

```
%: icgc-storage-client mount --manifest 4jdyyqa099ew22 --mount-point /tmp/
%: ls /tmp
```

Comprehensive User Guide at docs.icgc.org
Part 3.2. PCAWG Analytical Methods

PCAWG Workflows Available at Dockstore.org
Part 3.3. PCAWG Best Practices

Best Practices

• Look for PCAWG publications in 2017

• “Rogue’s Gallery of Cancer Genome Sequencing Artifacts” - examples of artifacts illustrated by rainfall plots, IGV, copy number profiles, BAM statistics, etc

• The code for detecting and filtering artifacts will also be made available