Analysis of RNA-Seq data at the Flemish Super Computer

Introduction to RNA-Seq Pipelines and Computational Analysis Mini Workshop

12 January 2018

Álvaro Cortés C.

GENOMICS CORELEUVEN

NGS Bioinformatics

NGS data means in general big data

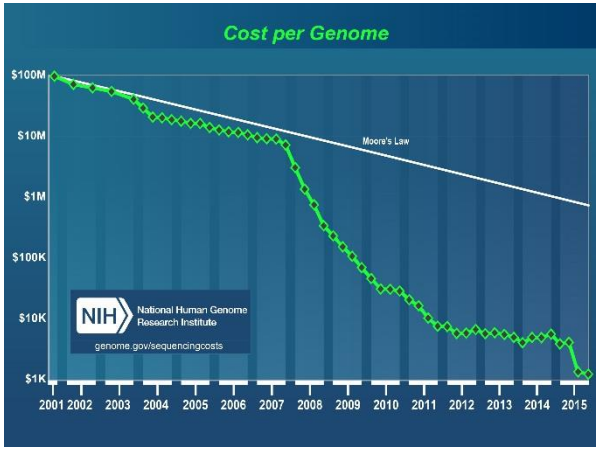
Raw data (uncompressed) per sample:

- Whole Human Genome:
  - 300Gb
- Exome data:
  - 6GB
- RNA-Seq
  - 1 GB

GENOMICS CORELEUVEN

NGS Data Analysis

Sequencing technology advances, analysis remains challenging



$1.000 sequencing and $10.000 bioinformatics costs?

GENOMICS CORELEUVEN

NGS Data Analysis

NGS data means big data...means big computing power

NGS data is usually analyzed on a supercomputer or cluster:

- UZ Leuven: Avalok/Hydra

- KU Leuven: VSC Flemish Super Computer

- Google genomics, etc.

GENOMICS CORELEUVEN

High Performance Computing (HPC)

- A computer cluster consists of multiple CPUs arranged in compute nodes
  - A typical compute node consists of 10 CPUs or cores and 128GB RAM
  - A cluster consists of multiple compute nodes
    - 208 nodes with two 10-core "Ivy Bridge" Xeon E5-2680v2 CPUs

GENOMICS CORELEUVEN

High Performance Computing: Taking Advantage

- To take full advantage of a computer cluster for genomics, computation preferably must be

  distributed across multiple cores and nodes

  - Otherwise, not necessarily more convenient than execution in desktop computer

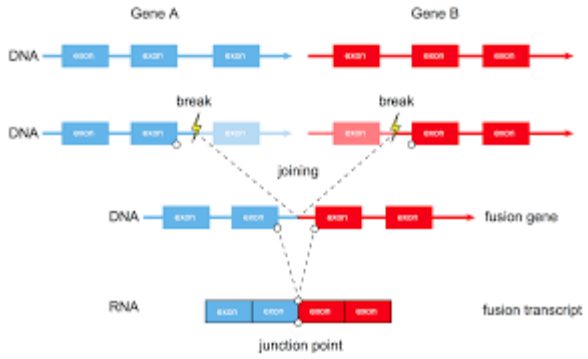
- Many NGS tools currently support multi-threading and/or multi-node computation: STAR, GATK 4.

GENOMICS CORELEUVEN

RNA-Seq and High Performance Computing

Is it needed?

- For differential expression, at the Genomics Core the analysis service is included for K.U. Leuven, UZ Leuven members
- Other pipelines: STAR, DeSeq2, Hisat2
- Analysis of public data
- Re-analysis with different experimental designs
- Downstream analysis: pathway analysis, gene ontology
- Other computational tasks:
  - Transcriptome assembly
  - Gene fusion analysis
  - SNP calling
  - Transcript discovery
  - Analysis non-model organisms

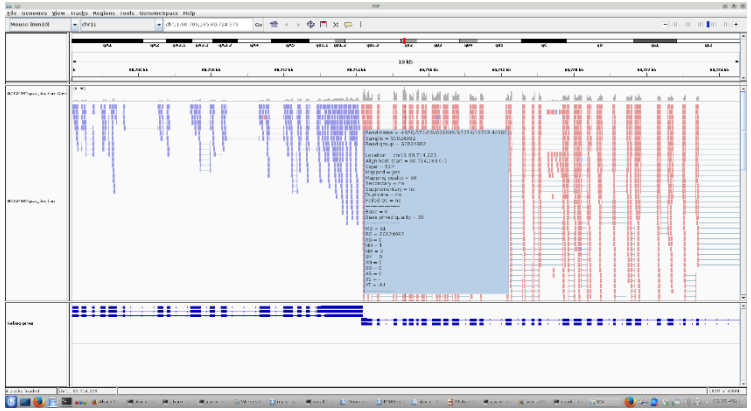


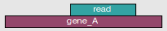
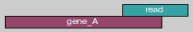
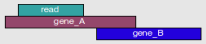
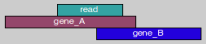
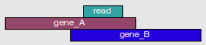
GENOMICS CORELEUVEN

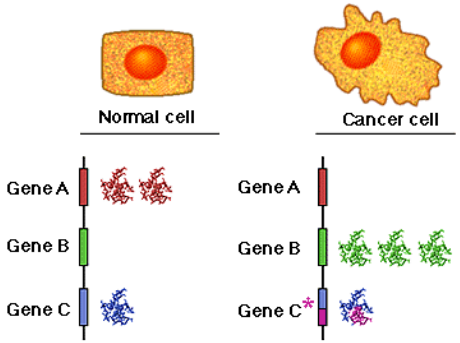
RNA-Seq and High Performance Computing

Differential Expression Analysis:

- Computationally intense for:
  - mapping,
  - counting,
  - and differential expression of genes



	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous



GENOMICS CORELEUVEN

RNA-Seq at VSC

To star working at the VSC:

- A VSC account and compute credits:

  https://www.vscentrum.be/

- Raw or preprocessed RNA-Seq data:

  Genomics Core transfers fastq files or preprocessed data (i.e. unnormalized counts) directly into a specified location at the VSC.

GENOMICS CORELEUVEN

RNA-Seq at VSC

Any analysis at the VSC:

- Number of nodes
- Number of cores per node
- Memory
- Maximum time the analysis will need the resources: walltime
  - Estimation?
- Project to charge the computation
- An email and a job name

PBS Script

GENOMICS CORELEUVEN

Portable Batch System (PBS)

Portable Batch System (or simply PBS) is the name of computer software that performs job scheduling. Its primary task is to allocate computational tasks, i.e., batch jobs, among the available computing resources. It is often used in conjunction with UNIX cluster environments. PBS is supported as a job scheduler mechanism by several meta schedulers ...

-- by Wikipedia

```
#!/bin/bash -l
#PBS -l walltime=48:00:00
#PBS -l mem=100gb
#PBS -l nodes=15:ppn=20
#PBS -M alvaro.cortes@uzleuven.be
#PBS -m aeb
#PBS -N star_mapping_population
#PBS -A lp_biogenomics
```

GENOMICS CORELEUVEN

RNA-Seq and PBS

- One PBS script per task:
  - Quality checking
  - Adaptor trimming
  - Mapping
  - Counting
  - Differential Expression analysis
    - EdgeR
    - DeSeq2
    - Bayseq
  - Summary of results and report

GENOMICS CORELEUVEN

RNA-Seq and more on PBS

- The input data
  - SAMPLE_DIR="$PROJECT_DIR/htseq";
- Location output
  - OUTPUT_DIR="$PROJECT_DIR/deseq2_16_10_2017_all"
- Software package to perform the actual analysis
  - Name
  - Version
    - module load R
- Input parameters software package:

```
Rscript $R_SCRIPT_DIR/deseq2.2conditions.R $CONDITION1
$CONDITION2 > $COMP_NAME".rscript.log" 2> $COMP\
_NAME".rscript.log";
```

GENOMICS CORELEUVEN

Submitting PBS Jobs

- Thinking

  $ qsub step8_deseq2.pbs

- Cerebro

  $ module load cerebro/2014a

  $ qsub step8_deseq2.pbs

- When a job is submitted, the job id is returned.

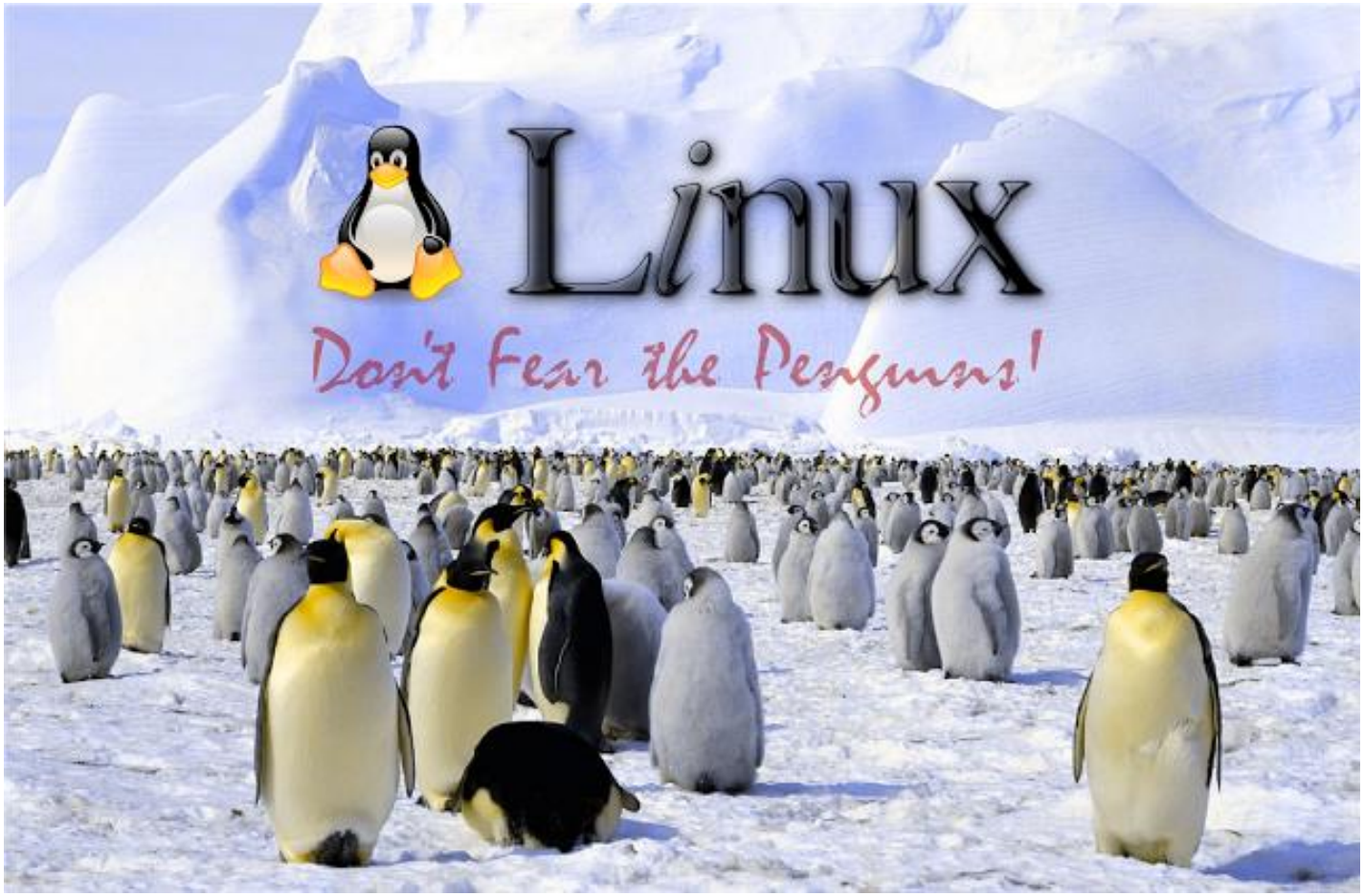
- jobIDs are unique.

  Thinking jobs starts with 2

  Cerebro jobs starts with 3

GENOMICS CORELEUVEN

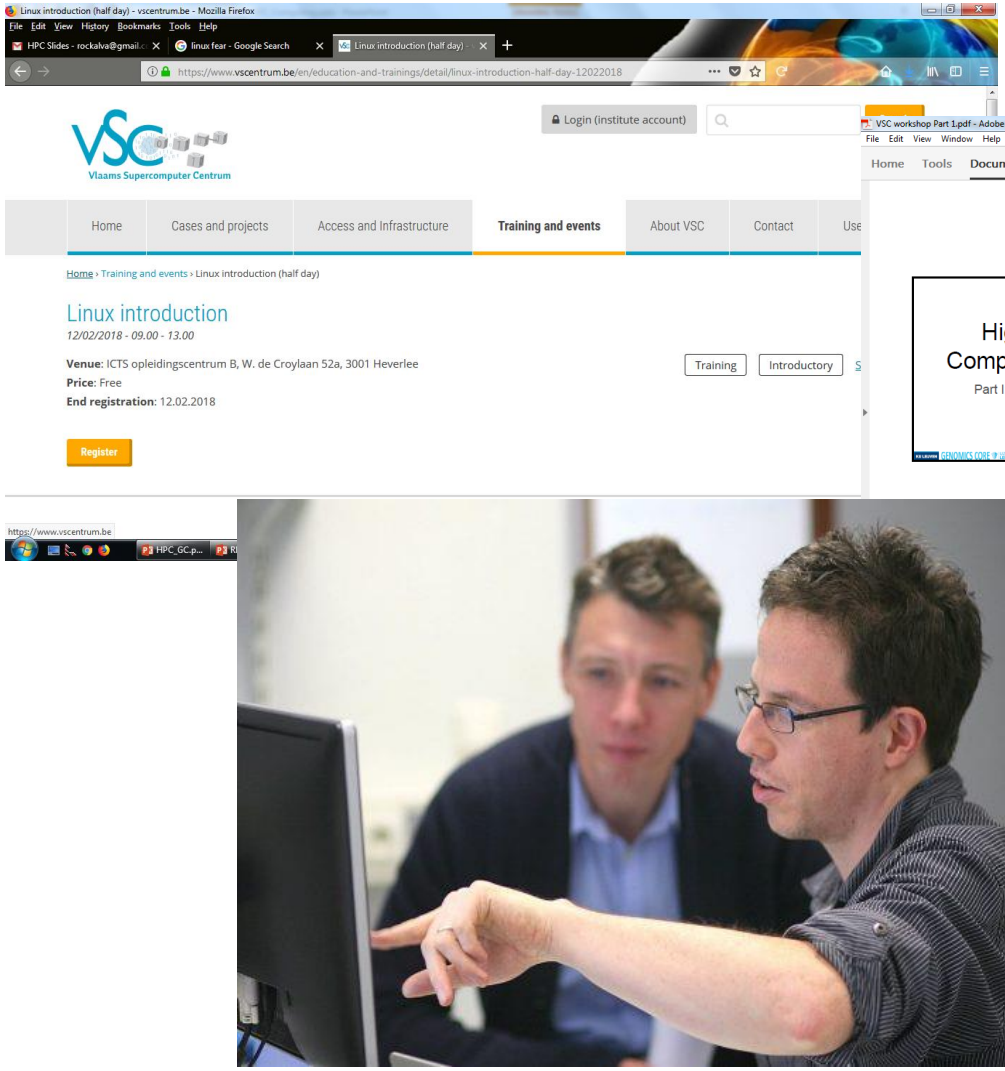
Hands-on Project



Linux - Don't Fear the Penguins by Rob A. Shinn

GENOMICS CORELEUVEN

An Ecosystem of Support



GENOMICS CORELEUVEN

LIVE Demo

GENOMICS CORELEUVEN

Thanks!

GENOMICS CORELEUVEN