

# Implementación de un sistema Big Data para el análisis de secuencias de ADN usando Apache Spark-ADAM

## Genómica Computacional

Quispe Gutierrez Lizeth<sup>1</sup>, Galvan Oyola Victor<sup>2</sup>, Cerron Tome Renzo<sup>3</sup>

<sup>123</sup>Computer Science  
National University of Engineering

Estadística, 01 Julio 2016



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

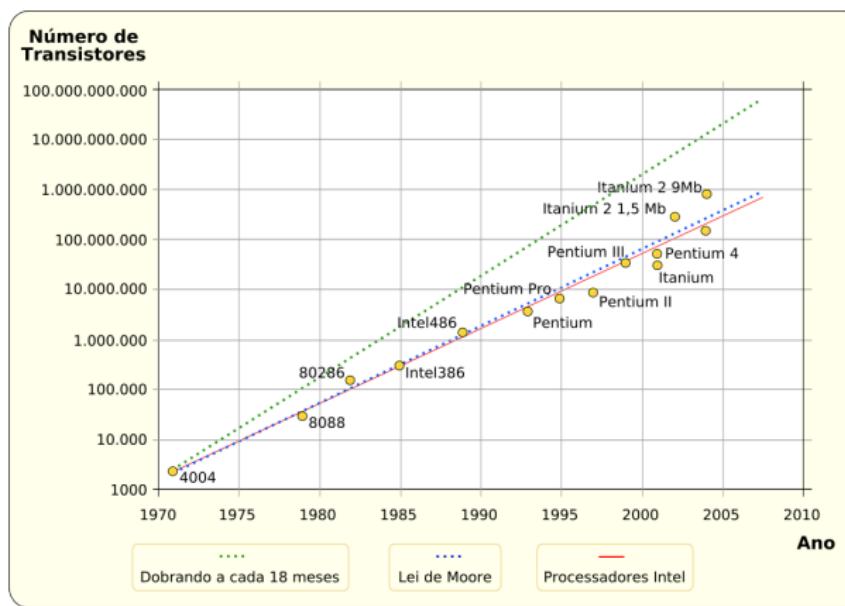
- Palindromic Sequence

## 5 Conclusiones



## Ley de Moore

"Cada 2 años el número de transistores en un microprocesador se duplica."



# Outline

## 1 Introducción

- Ley de Moore
- **Big Data**
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

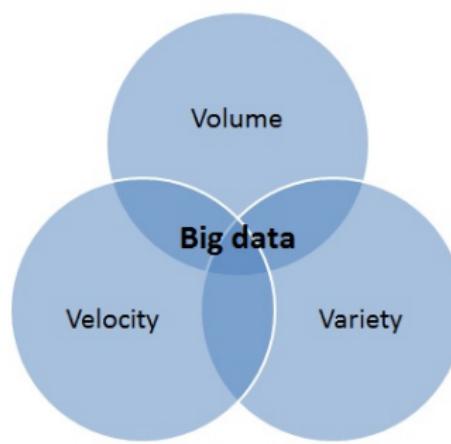
- Palindromic Sequence

## 5 Conclusiones



# Big Data

Big Data se define como el conjunto de herramientas informáticas destinadas a la manipulación, gestión y análisis de grandes volúmenes de datos de todo tipo los cuales no pueden ser gestionados por las herramientas informáticas tradicionales. Hace referencia al almacenamiento de grandes cantidades de datos.



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- **Big Data y Biología Computacional**
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



# Breve Descripción

Cada vez es mayor el volumen disponible de datos de origen biológico y éste abarca muchas áreas de la biología molecular, desde la ingente cantidad de información genética como resultado de experimentos de Next Generation Sequencing (NGS), hasta los datos de patrones de expresión determinados por la tecnología de DNA-microarrays



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- **Next Generation Sequencing**

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



# Definición de NGS

"Término general utilizado para describir una serie de diferentes tecnologías de secuenciación de ADN modernas."

"La cantidad de datos generados por métodos de Secuenciación de alto rendimiento (NGS) esta duplicándose cada 5 meses y la tendencia es que continuará por los siguientes años."



Figure: Plataforma de next generation sequencing(Illumina)



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



## Breve Descripción

"Apache Spark es un motor rápido y general para el procesamiento de grandes volúmenes de datos, con los módulos incorporados para streaming, SQL, aprendizaje automático y procesamiento gráfico."



## Breve Descripción

"Apache Spark es un motor rápido y general para el procesamiento de grandes volúmenes de datos, con los módulos incorporados para streaming, SQL, aprendizaje automático y procesamiento gráfico."

"Spark proporciona una interfaz para la programación de racimos enteros con paralelismo de datos implícita y tolerancia a fallos."



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones

# Ejemplos

ADAM es una plataforma genómica de análisis con formatos de archivos especializados construidos usando Apache Avro, Apache Spark y Apache Parquet.



# Ejemplos

ADAM es una plataforma genómica de análisis con formatos de archivos especializados construidos usando Apache Avro, Apache Spark y Apache Parquet.



# TimeLine

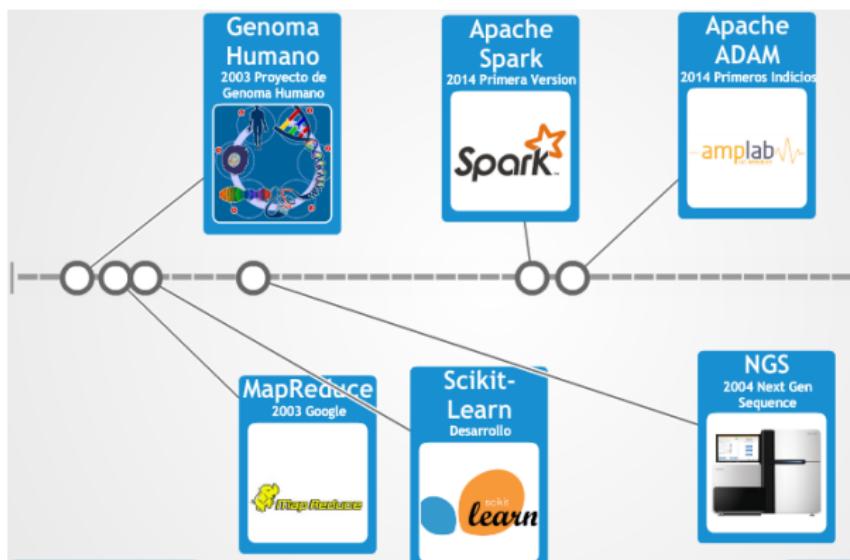


Figure: Línea de Tiempo de las tecnologías Big Data y Secuenciamiento de ADN



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN**

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



# Formatos

Algunos ejemplos de formatos de secuencias de ADN son los siguientes:

- SAM



# Formatos

Algunos ejemplos de formatos de secuencias de ADN son los siguientes:

- SAM Sequence Alignment Map



# Formatos

Algunos ejemplos de formatos de secuencias de ADN son los siguientes:

- SAM Sequence Alignment Map
- BAM



# Formatos

Algunos ejemplos de formatos de secuencias de ADN son los siguientes:

- SAM Sequence Alignment Map
- BAM Binary Sequence Alignment Map
- ADAM



# Formatos

Algunos ejemplos de formatos de secuencias de ADN son los siguientes:

- SAM Sequence Alignment Map
- BAM Binary Sequence Alignment Map
- ADAM Formato propuesta por el grupo bigdatagenomics y usada en el framework ADAM.



# Formatos

Algunos ejemplos de formatos de secuencias de ADN son los siguientes:

- SAM Sequence Alignment Map
- BAM Binary Sequence Alignment Map
- ADAM Formato propuesta por el grupo bigdatagenomics y usada en el framework ADAM.

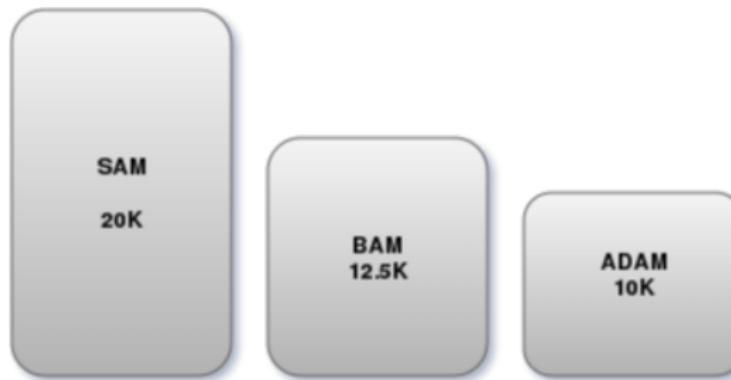


Figure: Comparación de tamaños de formatos para secuencias de ADN

# Sátira

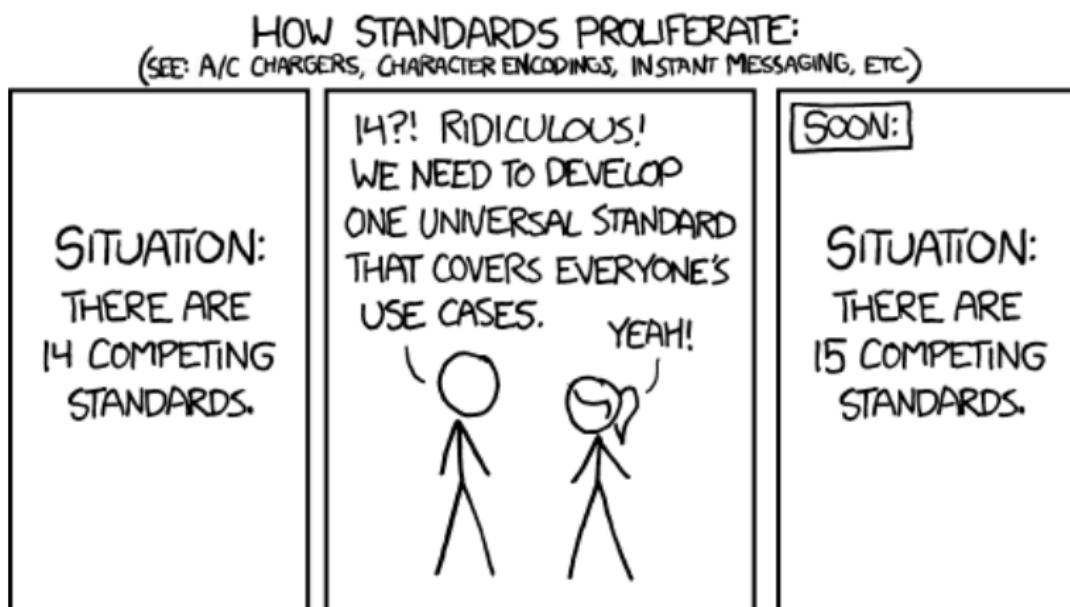


Figure: Sátira de standards en general.



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



# Pruebas con ADAM en el cluster de máquinas virtuales

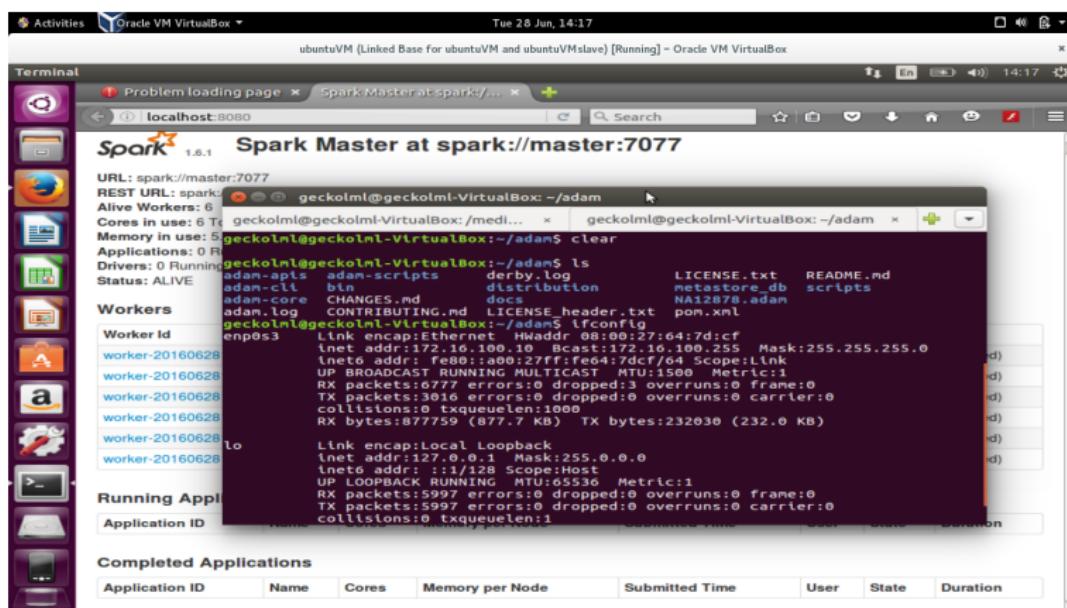


Figure: Sátira de standards en general.



## Cluster con máquinas virtuales

## Pruebas con ADAM en el cluster de máquinas virtuales

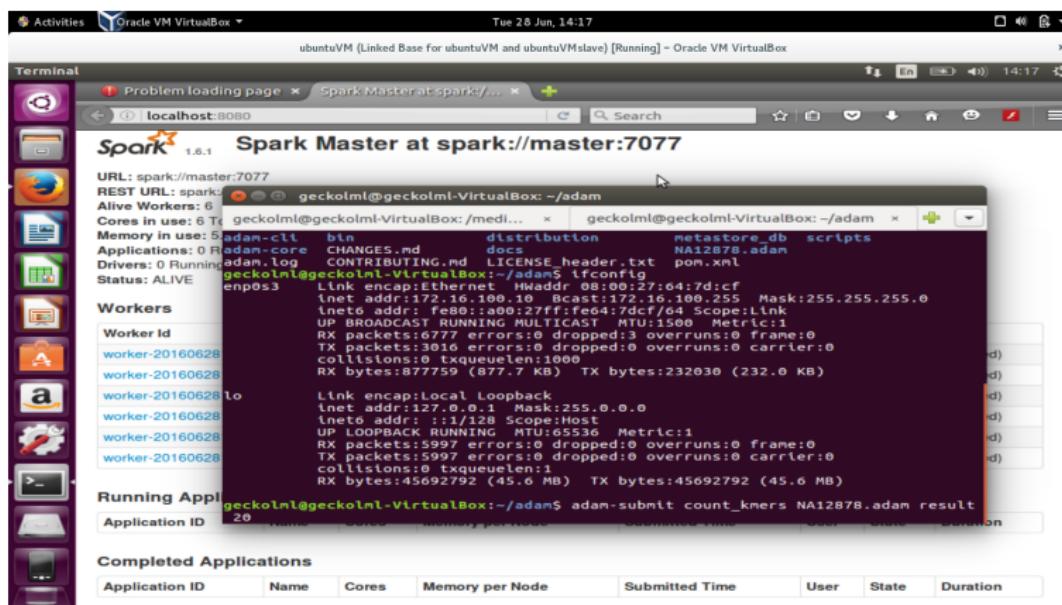


Figure: Ejecución del programa count\_kmers en ADAM para el archivo entrada NA12878.adam



### Cluster con máquinas virtuales

## Pruebas con ADAM en el cluster de máquinas virtuales

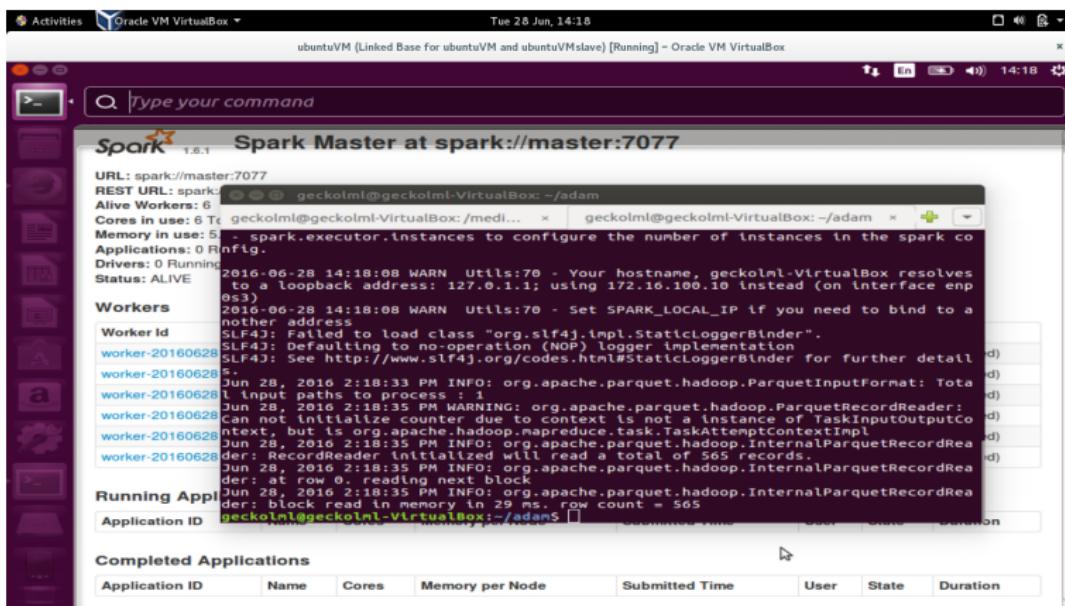


Figure: Registro luego de haber corrido el programa count\_kmers



# Pruebas con ADAM en el cluster de máquinas virtuales

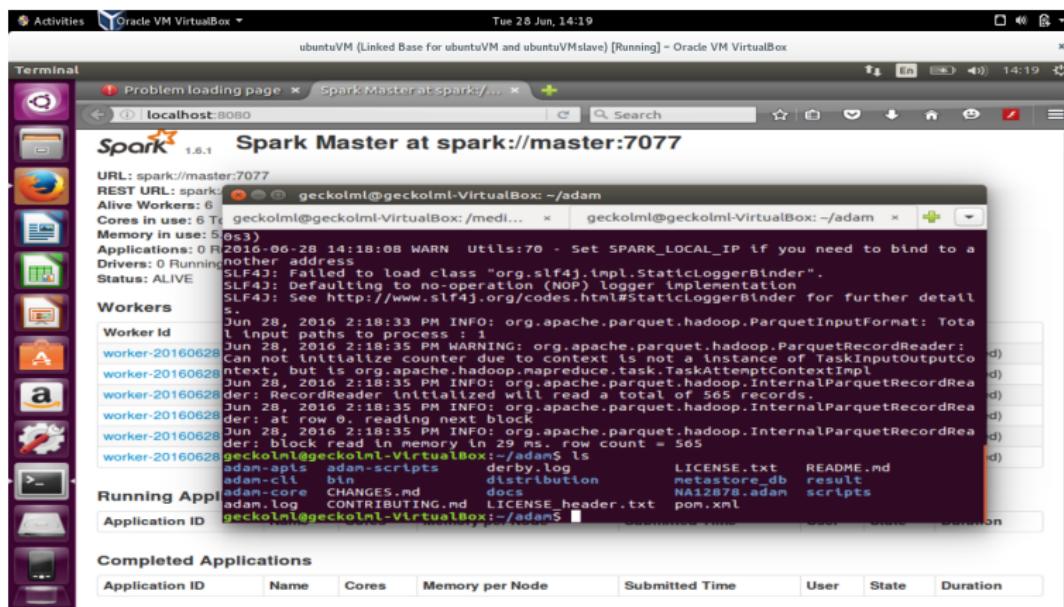


Figure: Archivos actuales luego de correr count\_kmers para k=20



# Pruebas con ADAM en el cluster de máquinas virtuales

The screenshot shows a terminal window titled "Activities" with a tab for "Oracle VM VirtualBox". The main area displays the "Spark Master at spark://master:7077" interface. The URL is `spark://master:7077`. The master has 6 cores in use and 0 running applications. It lists 6 workers, each with a unique ID (e.g., worker-20160628) and their respective memory usage. Below the workers, it shows 3 completed applications. The bottom section is a table for "Completed Applications" with columns: Application ID, Name, Cores, Memory per Node, Submitted Time, User, State, and Duration.

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
1	count_kmers	1	1	2016-06-28 14:19:29	geckolml	FINISHED	00:00:00.000
2	count_kmers	1	1	2016-06-28 14:19:29	geckolml	FINISHED	00:00:00.000
3	count_kmers	1	1	2016-06-28 14:19:29	geckolml	FINISHED	00:00:00.000

Figure: Archivo resultado de count\_kmers para k 20



# Pruebas con ADAM en el cluster de máquinas virtuales

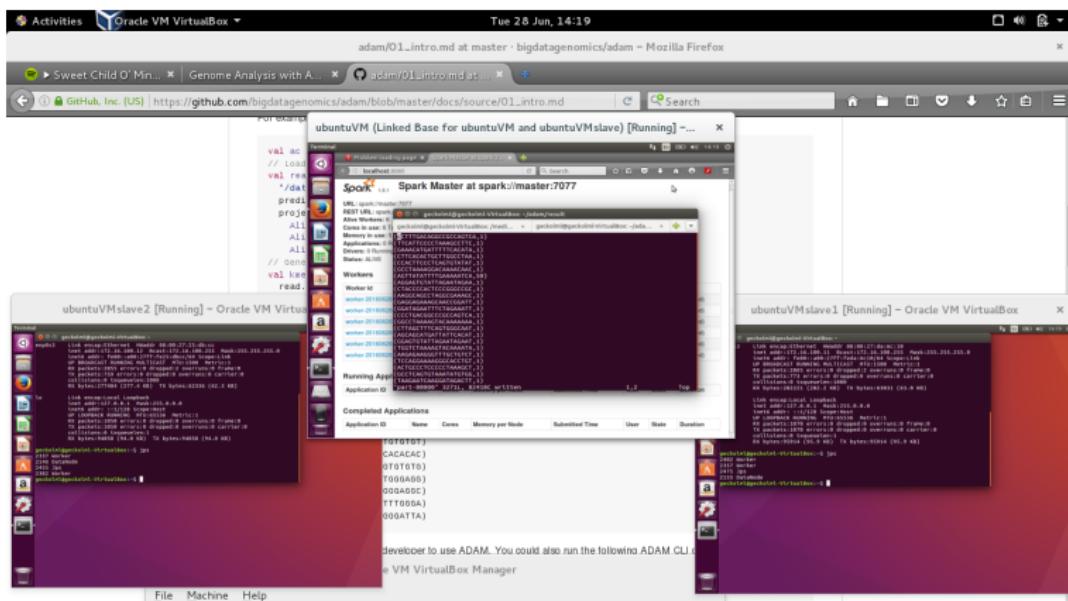


Figure: Vista de todas las máquinas virtuales que conforman el cluster



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



# Video



# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- **Explicación de count\_kmers**

## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



Explicación de count\_kmers

# Código fuente de count\_kmers

```
package org.bdgenomics.adam.cli

import java.util.logging.Level
import org.apache.hadoop.mapreduce.Job
import org.apache.spark.SparkContext
import org.apache.spark.rdd.RDD
import org.bdgenomics.adam.projections.{ AlignmentRecordField, Projection }
import org.bdgenomics.adam.rdd.ADAMContext._
import org.bdgenomics.adam.util.ParquetLogger
import org.bdgenomics.formats.avro.AlignmentRecord
import org.bdgenomics.utils.cli._
import org.bdgenomics.utils.misc.Logging
import org.kohsuke.args4j.{ Argument, Option => Args4jOption }

object CountReadKmers extends BDGCommandCompanion {
    val commandName = "count_kmers"
    val commandDescription = "Counts the k-mers/q-mers from a read dataset."

    def apply(cmdLine: Array[String]) = {
        new CountReadKmers(Args4j[CountReadKmersArgs](cmdLine))
    }
}
```

## Explicación de count\_kmers

## Código fuente de count\_kmers

```
class CountReadKmersArgs extends Args4jBase with ParquetArgs {
    @Argument(required = true, metaVar = "INPUT", usage = "The ADAM, BAM or SAM file to count kmers from", index = 0)
    var inputPath: String = null
    @Argument(required = true, metaVar = "OUTPUT", usage = "Location for storing k-mer counts", index = 1)
    var outputPath: String = null
    @Argument(required = true, metaVar = "KMER_LENGTH", usage = "Length of k-mers", index = 2)
    var kmerLength: Int = 0
    @Args4jOption(required = false, name = "-print_histogram", usage = "Prints a histogram of counts.")
    var printHistogram: Boolean = false
    @Args4jOption(required = false, name = "-repartition", usage = "Set the number of partitions to map data to")
    var repartition: Int = -1
}
```



## Explicación de count\_kmers

## Código fuente de count\_kmers

```

class CountReadKmers(protected val args: CountReadKmersArgs) extends BDGSparkCommand[CountReadKmersArgs] with Logging {
    val companion = CountReadKmers

    def run(sc: SparkContext) {

        // Quiet Parquet...
        ParquetLogger.hadoopLoggerLevel(Level.SEVERE)

        // read from disk
        var adamRecords: RDD[AlignmentRecord] = sc.loadAlignments(
            args.inputPath,
            projection = Some(Projection(AlignmentRecordField.sequence))
        )

        if (args.repartition != -1) {
            log.info("Repartitioning reads to '{} partitions".format(args.repartition))
            adamRecords = adamRecords.repartition(args.repartition)
        }

        // count kmers
        val countedKmers = adamRecords.countKmers(args.kmerLength)

        // cache counted kmers
        countedKmers.cache()

        // print histogram, if requested
        if (args.printHistogram) {
            countedKmers.map(kv => kv._2.toLong)
                .countByValue()
                .toSeq
                .sortBy(kv => kv._1)
                .foreach(println)
        }

        // save as text file
        countedKmers.saveAsTextFile(args.outputPath)
    }
}

```



Figure: Código fuente de prueba CountReadKmers.scala del proyecto ADAM

# Outline

## 1 Introducción

- Ley de Moore
- Big Data
- Big Data y Biología Computacional
- Next Generation Sequencing

## 2 Big Data en secuencias de ADN

- Apache Spark
- Apache ADAM (Spark, Avro, Parquet )
- Formatos de secuencias de ADN

## 3 Desarrollo de la Propuesta

- Cluster con máquinas virtuales
- Video
- Explicación de count\_kmers

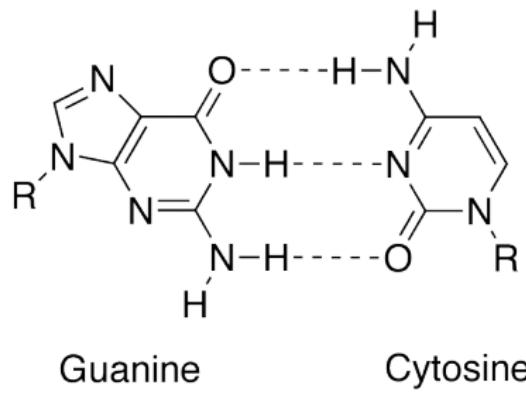
## 4 Otras aplicaciones en Genómica Computacional

- Palindromic Sequence

## 5 Conclusiones



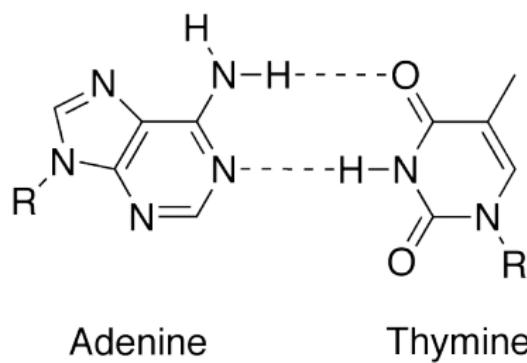
## Complementariedad de ADN



**Figure:** Coincidencia entre bases de ADN(Citosina y Guanina) unidas con enlaces de hidrógeno



## Complementariedad de ADN



**Figure:** Coincidencia entre bases de ADN(Adenina y Timina) unidas con enlaces de hidrógeno

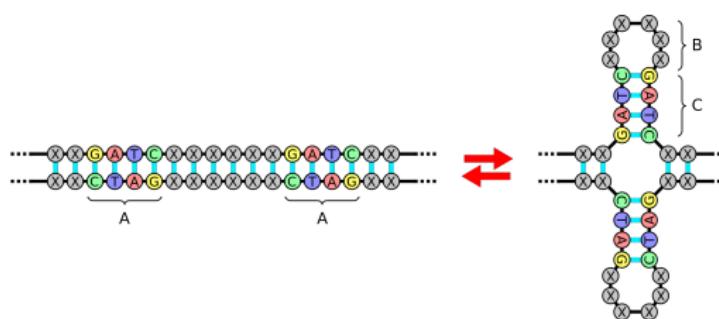


# Watson-Crick Palindrome

Se dice que una secuencia de ADN es Watson-Crick Palindrome si una secuencia de ADN es igual a su reverso complementario.

Ejemplo:

*AGCTATGATCATAGCT*



**Figure:** Repetición invertida A) secuencia de repetición invertida, B) secuencia de repetición directa, C) Tallo con el apareamiento de bases de las secuencias repetidas invertidas



## Conclusion

Concluimos lo siguiente:

- La tasa de crecimiento de la generación de datos de secuencia está superando la de capacidades de hardware(Ley de Moore) en un factor de aproximadamente cuatro de cada año.



## Conclusion

Concluimos lo siguiente:

- La tasa de crecimiento de la generación de datos de secuencia está superando la de capacidades de hardware(Ley de Moore) en un factor de aproximadamente cuatro de cada año.
  - Tecnologías Big Data proporciona una herramienta al investigador para poder almacenar, recuperar y analizar una gran cantidad de datos.



# Conclusion

Concluimos lo siguiente:

- La tasa de crecimiento de la generación de datos de secuencia está superando la de capacidades de hardware(Ley de Moore) en un factor de aproximadamente cuatro de cada año.
- Tecnologías Big Data proporciona una herramienta al investigador para poder almacenar, recuperar y analizar una gran cantidad de datos.
- La Biología moderna necesita de fuertes herramientas computacionales para contrarrestar la inmensa cantidad de datos que desea analizar( NGS, Genoma Humano, Cytometry).



# Conclusion

Concluimos lo siguiente:

- La tasa de crecimiento de la generación de datos de secuencia está superando la de capacidades de hardware(Ley de Moore) en un factor de aproximadamente cuatro de cada año.
- Tecnologías Big Data proporciona una herramienta al investigador para poder almacenar, recuperar y analizar una gran cantidad de datos.
- La Biología moderna necesita de fuertes herramientas computacionales para contrarrestar la inmensa cantidad de datos que desea analizar( NGS, Genoma Humano, Cytometry).
- La llegada de GPUS(Graphics Processor Unit) a transformado nuestra investigación en biología mejorando los tiempos de análisis de los datos flujo de citometría, mejorando el desarrollo de varios proyectos para hacer frente al cancer.



# Conclusion

Concluimos lo siguiente:

- La tasa de crecimiento de la generación de datos de secuencia está superando la de capacidades de hardware(Ley de Moore) en un factor de aproximadamente cuatro de cada año.
- Tecnologías Big Data proporciona una herramienta al investigador para poder almacenar, recuperar y analizar una gran cantidad de datos.
- La Biología moderna necesita de fuertes herramientas computacionales para contrarrestar la inmensa cantidad de datos que desea analizar( NGS, Genoma Humano, Cytometry).
- La llegada de GPUS(Graphics Processor Unit) a transformado nuestra investigación en biología mejorando los tiempos de análisis de los datos flujo de citometría, mejorando el desarrollo de varios proyectos para hacer frente al cancer.



# Bibliografia I

📘 Sandy Ryza, Uri Laserson, Sean Owen & Josh Wills  
*Advanced Analytics with Spark.*

Patterns for Learning from data at scale

📘 Eric S. Lander  
*Initial impact of the sequencing of the human genome.*  
Review

📘 Cliburn Chan  
*Big Data in Computational Biology*  
An invitation to the digital science of life

📘 Big Data Genomics  
<http://bdgenomics.org>  
Archives

