

# Analysis of Whole Genom

Jarlier

(1) Institut Curie, Paris France, (2) INSERM U900, Paris

## Motivation:

The Next Generation Sequencing (NGS) technology offers new insights in cancer research. The decreasing cost of sequencing, whole genome sequencing becomes more widely used in research. The amount of data generated is increasing and also the complexity of the downstream analysis. For instance, a whole genome sequencing pipeline lacks of scalability.

To tackle traditional bottlenecks, we have used a parallelization with message passing interface (MPI). We also provides many optimizations such as collective operations to optimize IO ; it provides

The classical NGS pipeline consists of two steps. The first step is the alignment of small fragments to a reference genome and genomic position. These operations are essential but time consuming. Therefore we have implemented and then we show the results we have obtained on whole human genome sequencing.

## 1) Description of the MPI workflow

Every NGS pipeline starts with the two following operations: the alignment and the sorting. The position is called the coverage of a sample. The deeper the coverage the better the results. A 30x coverage produces 1 TB of data.

After intensive study of the alignment and sorting algorithms we have noticed that they are not optimized at network level. Other optimizations have been implemented such as : collective operations

.fastq

```
TTTACGA
ACGACT
TTTACGA
TCCTAGC
...
GCTGCTA
AGCTGCC
AGCTGCC
```

Step 2 :  
- Alignment (BWA MEM)  
- Shared memory

## Overview of the alignment algorithm

## 2) Description of the IT architecture

For fast reading and writing, collective operations are mandatory. They can be achieved upon a custom distributed file system. At Institut Curie we have tested our solution on Lustre with the same performances.

computing

