From Exome to Whole Genome sequencing analysis

Florian SANDRON[1,3], Lilia MESROB [1,2], Nicolas WIART [1], Aurélie LEDUC [1], Ghislain SEPTIER [1], Stephane MESLAGE [1], Delphine BACQ [1], Vincent MEYER [1], François ARTIGUENAVE [1]

[1]CEA/LBI (Laboratoire de Bioinformatique et Informatique) ,[2] INSERM, [3] INRA/France Génomique
Centre National de GénotypageCommissariat à l'énergie Atomique et aux Énergies Alternatives
Contact: *vincent.meyer@cng.fr*

1. Varscope: High throughput calibrated process

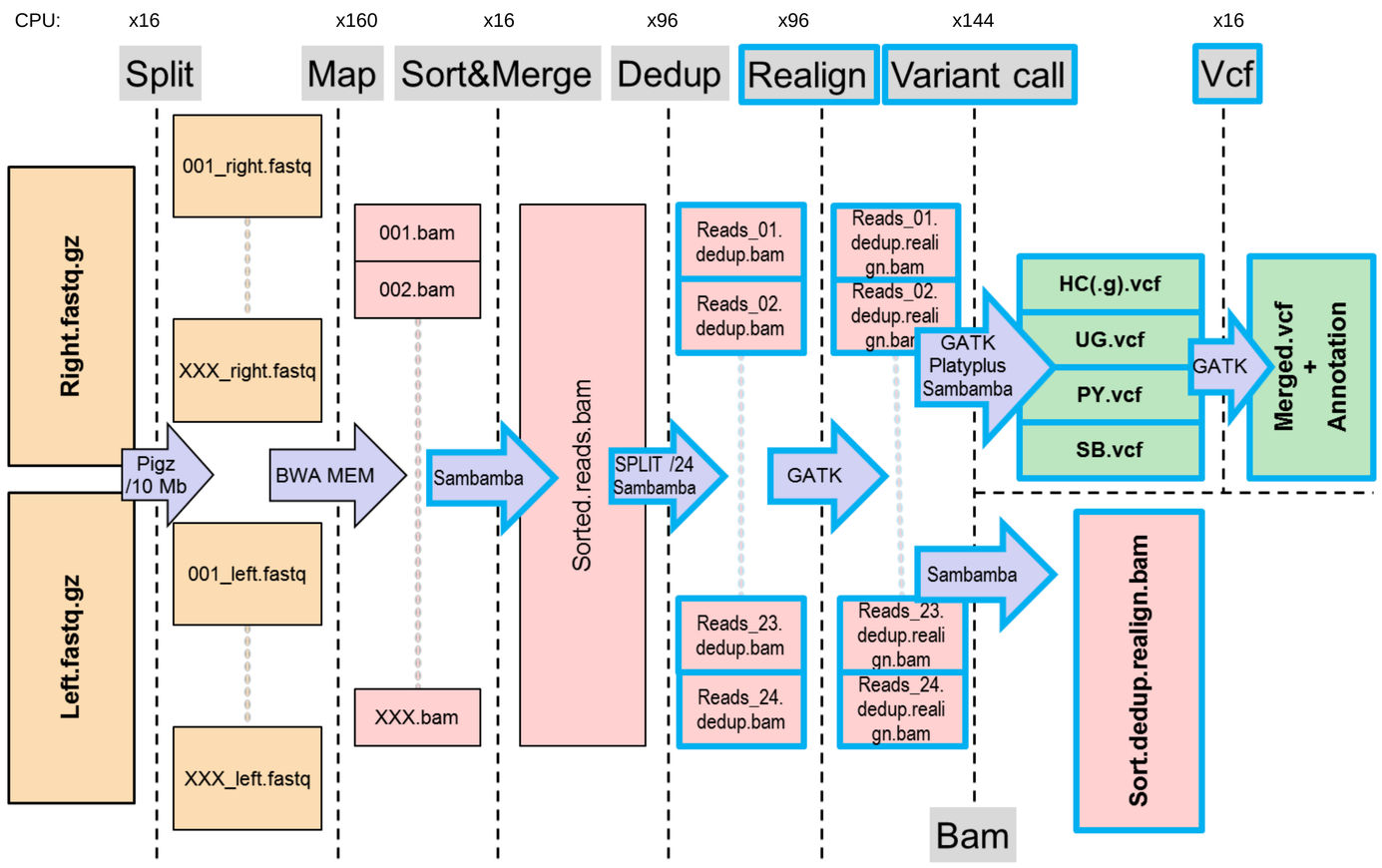
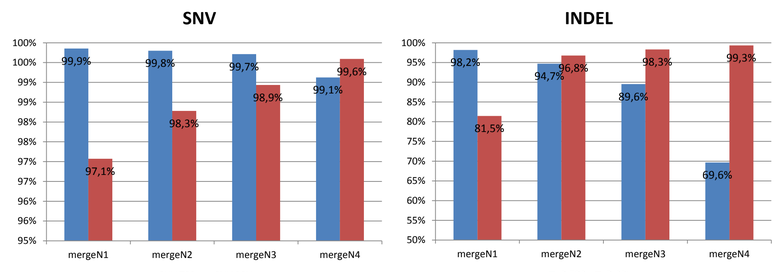


Fig. 1: Overview of the CNG/CCRT mapping and variant calling pipeline organization. Last developments are highlighted in blue

	#base	Fastq.gz	bam	Vcf	#variant
Exome 70x	8,7 Gb	9,7 Go	5,7 Go	16 Mo	75 483
WG 30x	98 Gb	84 Go	75 Go	1075 Mo	4 792 544
WG 70x	223 Gb	196 Go	187 Go	1166 Mo	4 958 931

Fig. 2: Sizes/Values summary related to the files generated during WES and GWS analysis

- **Optimization of the« Map-reduce » strategy (fig.1)**
- **Softwares upgrade and update (fig.1)**
- **Addition of a « multicall» strategy (fig1,3)**

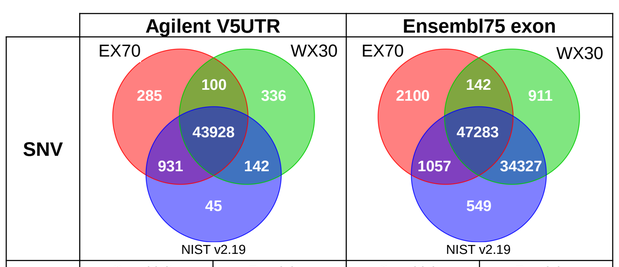


40x Depth	HC	UG	Platypus	Sambamba	N1	N2	N3	N4
SNV TPR	99,71%	99,69%	99,68%	99,40%	99,85%	99,80%	99,71%	99,12%
SNV PPV	98,93%	98,11%	98,09%	98,95%	97,07%	98,28%	98,94%	99,59%
InDel TPR	95,62%	88,33%	92,80%	72,72%	98,20%	94,69%	89,55%	69,64%
InDel PPV	96,20%	98,03%	93,63%	82,04%	81,45%	96,77%	98,31%	99,33%
F1 SNV	99,32%	98,90%	98,88%	99,18%	98,44%	99,03%	99,32%	99,36%
F2 SNV	99,55%	99,37%	99,36%	99,31%	99,28%	99,49%	99,56%	99,22%
F1 ID	95,91%	92,93%	93,21%	77,10%	89,04%	95,72%	93,73%	81,88%
F2 ID	95,73%	90,11%	92,97%	74,42%	94,32%	95,10%	91,18%	74,07%

Platypus	UnifiedGenotyper	SamBamba	SamTools	HaplotypeCaller
5 / 360	29 / 2088	37 / 2664	118 / 8496	254 / 18288

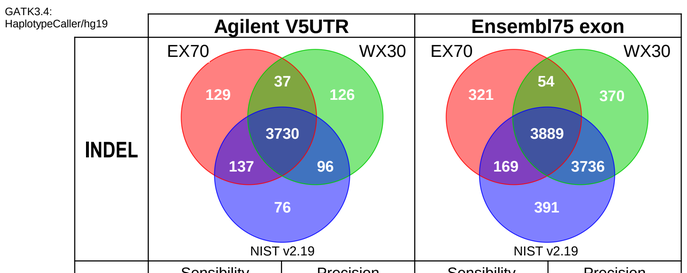
Fig. 3: Above calling programm and « multicall » strategy performance **(NIST/GIAB.v2.19)** and approximate Real-Time/Machine-Time(min/72cores)

3. Exome to Whole Genome transition



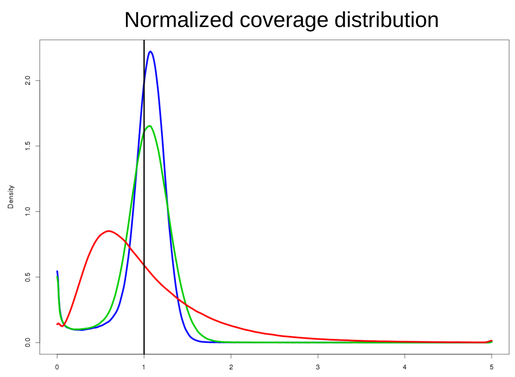
	Sensitivity	Precision	Sensitivity	Precision
Exome (70x)	99,58 %	99,15 %	58,09 %	95,57 %
Genome (30X)	97,83 %	99,02 %	98,07 %	98,73 %
Genome (70X)	98,84 %	99 %	98,92 %	98,64 %

GATK3.4: HaplotypeCaller/hg19



	Sensibility	Precision	Sensibility	Precision
Exome (70x)	95,74 %	95,88 %	49,58 %	91,54 %
Genome (30X)	94,73 %	95,91 %	93,16 %	94,73 %
Genome (70X)	96,98 %	96,29 %	95,76 %	94,63 %

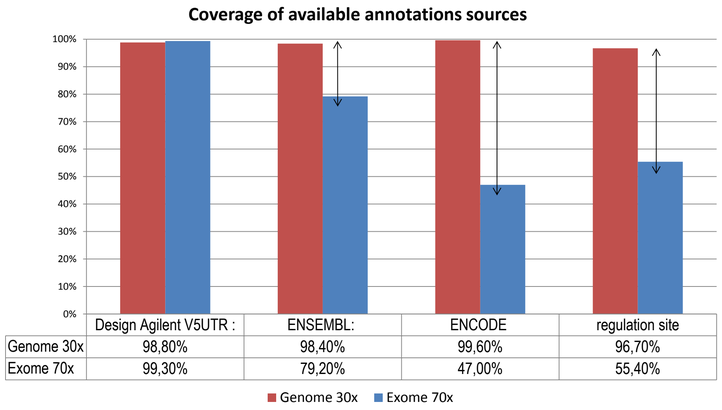
Fig. 7: Variant calling comparison between WES and WGS on Agilent V5UTR exome targeted sequencing and Ensembl « exons » selection(NIST/GIAB)



Exome 70X: dispersion = 0.479974
Genome 30X: dispersion = 0.196631
Genome 70X: dispersion = 0.173464

$$dispersion = \frac{\sigma^2 - \mu}{\mu^2}$$

Fig. 8: Comparison of Coverage dispersion between WES and WGS on AgilentV5-UTR targeted regions



	Design Agilent V5UTR :	ENSEMBL:	ENCODE	regulation site
Genome 30x	98,80%	98,40%	99,60%	96,70%
Exome 70x	99,30%	79,20%	47,00%	55,40%

Fig. 9: Annotation coverage comparision between WES and WGS according to Ensembl 75/GRCh37.p13

- **WGS Calling perfomance mostly similar on exome targeted regions (fig.7)**
- **Steadier coverage for WGS (fig.8)**
- **Wider annotation access (fig7,9)**

Abstract

The last few years show an enthusiastic evolution of the human genome resequencing projects profile. We reached the point where Whole Genome Sequencing (WGS) projects are about to become a new standard for research and personal genomics. To achieve this goal, major challenges have to be addressed: Management of massive data storage (1Po/year for X5 illumina), Optimization of computing resources and development of new analysis strategies. Here we present a focus on the last progress carried out at the Centre National de Génotypage to handle this leading transition phase.

Since 2013, bioinformatics analyses of the CNG are processed at the HPC facility *TGCC: CEA, "Très Grand Centre de Calcul", Bruyère le Chatel*, 2 Pflops including 400 Tflops and 5 PBytes dedicated to France Genomics projects. For example, pipelines for exome sequencing data analysis (varscope), mainly based on best practices for mapping, calling and first pass annotation (***DePristo et al., 2011/Van der Auwera et al., 2013***) was setup in 2013 but required a major upgrade to ensure speed and scalability fitting the production flow anticipation (up to 9000 WG by year for a X5 illumina sequencer).

We present here the high throughput calibrated process deployed on the TGCC facility (fig.1). This process include a “map-reduce” optimization and software upgrade to handle the 10 times increase capacity required by the transition of Whole Exome Sequencing (WES) to WGS (fig.2).

Moreover, we benchmarked several variant calling algorithms (***NIST/GIAB; [1]Zook et al., 2014***) and added “4-multicall” step to our process based on HaplotypeCaller(HC; ***[2]McKenna et al.,2010***), UnifyedGenotyper(UG), Platypus(PY; ***[3]Rimmer et al., 2014***), and Sambamba/Samtools(SB/ST; ***[4]Tarasov et al., 2015; [5]Li et al., 2009***)(fig.1, fig. 3).

We evaluated our process from low coverage analysis (10X, ex: very large cohort sequencing), to standard coverage (30X) and finally high coverage sequencing (> 100x, ex: somatic mutation or mosaicism detection). In “standard” condition, data from a 30X experiment are currently completed in approx. 8 hours (fig. 1). We didn’t identify any deadlock situation up to a 240X coverage (equivalent to a full X5 flowcell). Finally, we compared relative performance of WES and WGS and showed equivalent coding variants detection, steadier genome coverage distribution and up to 50% wider access to biologically relevant annotated regions (fig 7,8,9).

In conclusion, we are facing a major transition which requires refactoring and development to harvest knowledge from constantly over-flooding sources of data. We showed here benefits brought by WGS over WES in term of coverage and precision and important upgrades realized to support large WGS programs.

Our next step in this context will be to evaluate and add compression data strategies for storage and processing optimization. In addition we will add our structural variation process to the production flow to provide our collaborators a more complete view of the genome organization.

2. Scalability: from low to high coverage

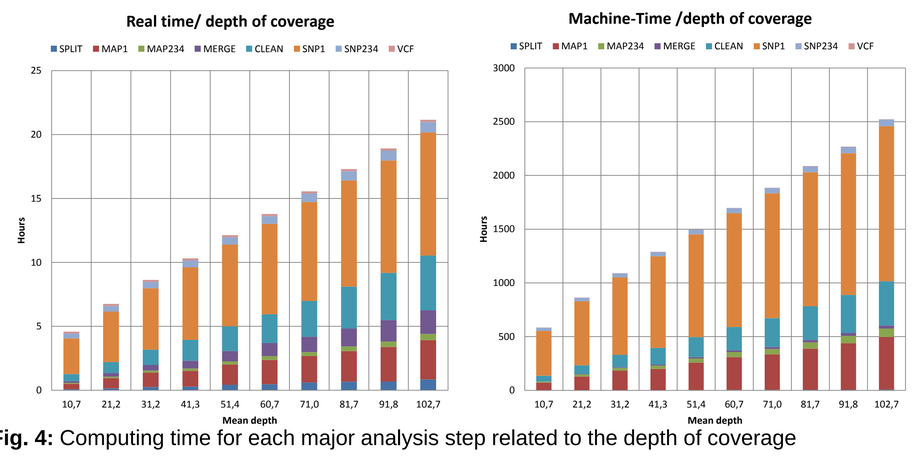


Fig. 4: Computing time for each major analysis step related to the depth of coverage

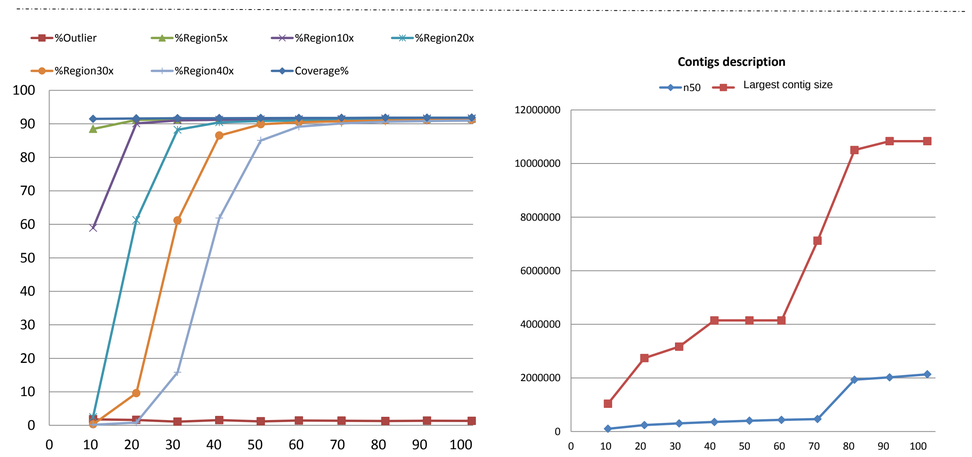


Fig. 5: Coverage metrics distribution related to the depth of coverage

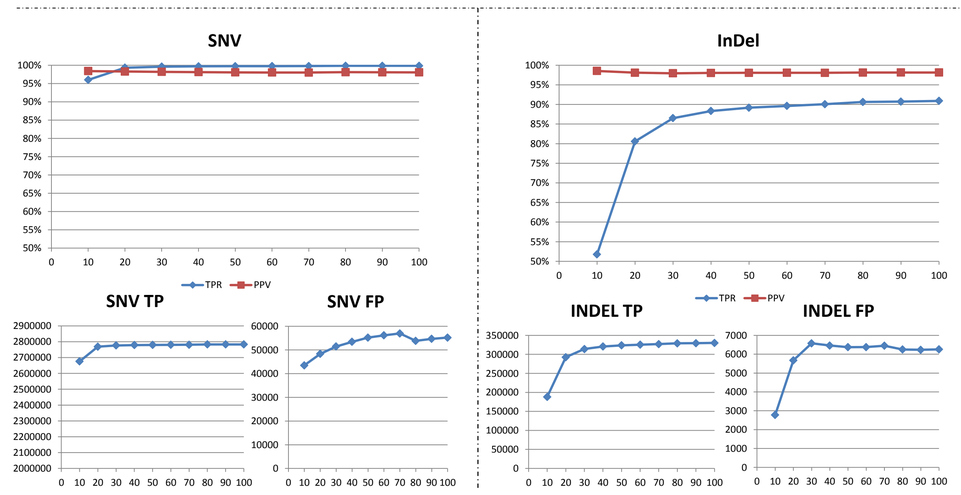


Fig. 6: Evolution of the indel/SNV sensitivity/precision related to the depth of coverage (NIST/GIABv2.19)

- **2,5h for a WES 60X**
- **8h for a WGS 30x up to 42h for a WGS 240x**
- **100 WES (70X) / ~12 WGS (30X) overnight**

Perspectives

- **Addition of compression data strategies (cf. JY.Lalanne; N.Wiart)**
- **Switch to GRCh38 as default analysis environment**
- **Addition of our structural variation detection process to the production flow**