



VNIVERSITAT
DE VALÈNCIA

bioaraba

osasun ikerketa institutua
instituto de investigación sanitaria

INCLIVA | VLC
Biomedical Research Institute

Carolina Monzó Cataluña

Guionar Pérez de Nanclares

Ana Bárbara García García

Vicente Arnau Llombart



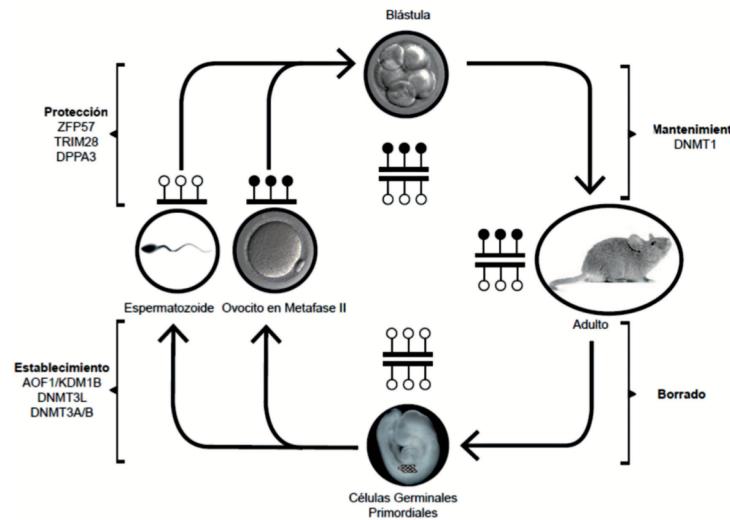
Estudio de alteraciones
genéticas en pacientes
con enfermedades de
impronta

1. Introducción

Impronta genómica

Marcas epigenéticas hereditarias determinadas por el origen parental.

- DMRs maternas: islas CpG en promotores y factores de transcripción.
- DMRs paternas: islas CpG en zonas intergénicas.



Enfermedades de impronta

PHP

Pseudohipoparatiroidismo

- Resistencia a hormona PTH.
- Talla baja, obesidad.
- Actividad reducida de la proteína G estimuladora.

BWS

Beckwith-Wiedemann

- Trastorno del crecimiento, visceromegalia.
- Hipoglucemia.

SRS

Silver-Russell

- Retraso en el crecimiento.
- Macrocefalia, micrognatia, dificultad de alimentación.
- Retraso en el desarrollo psicomotor.

ANS

Angelman

- Discapacidad intelectual grave.
- Epilepsia, trastornos del movimiento.
- Microcefalia, alteraciones del lenguaje.

TS

Temple

- Retraso en el crecimiento.
- Hipotonía, obesidad.
- Pubertad de inicio temprano.

MLID

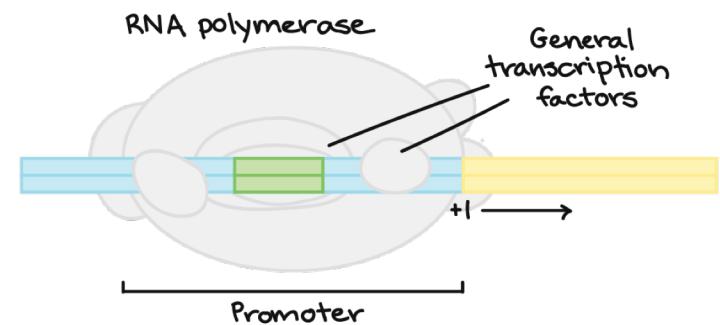
Multilocus

- Comorbilidades.

Enfermedades de impronta

- Mutaciones en el alelo no metilado.
- Translocaciones.
- CNVs.
- Disomía uniparental.

- Defectos de impronta.
 - *Cis*
 - *Trans*



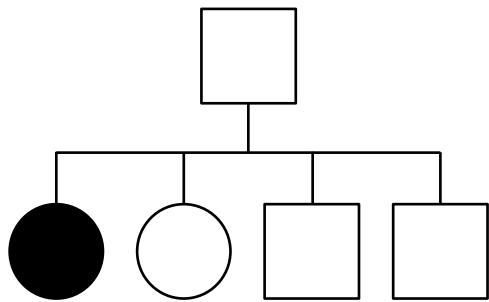
2. Objetivos

- Analizar patrones de herencia de variantes, para lo cual se desarrollará un método bioinformático para realizar el análisis de la secuenciación de Exoma completo y el estudio de familias.
- Estudiar datos de secuenciación de alta profundidad (*Deep-Seq*), para ello, se desarrollará un segundo método bioinformático dirigido al estudio de variantes exónicas, intrónicas, intergénicas, secuencias repetitivas y CNVs.
- Validar las metodologías, comparando los datos obtenidos utilizando las *pipelines* desarrolladas en este trabajo con resultados de estudios con técnicas ortogonales y datos “*Gold Standard*” publicados.
- Analizar las muestras en estudio en busca de alteraciones asociadas a los fenotipos de los pacientes y genes de interés que puedan estar actuando en *trans* en la regulación de la metilación en MLID.

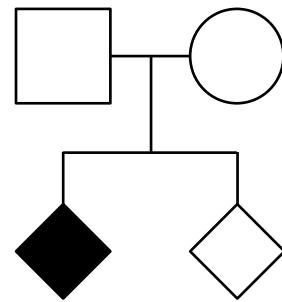
3. Material y Métodos

Familias estudiadas mediante secuenciación de Exoma

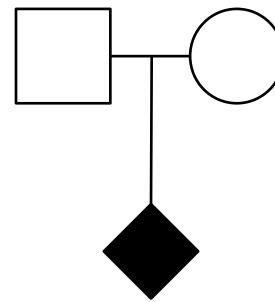
1 Quinteto



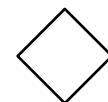
3 Cuartetos



6 Tríos



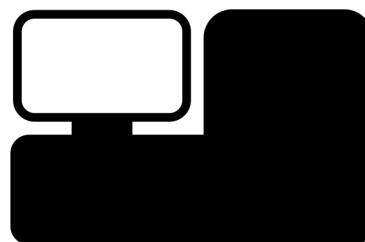
7 Individuales



- 10 pacientes afectados por PHP.
- Estudios previos: MS-MLPA y array de metilación.
- 35 familiares.

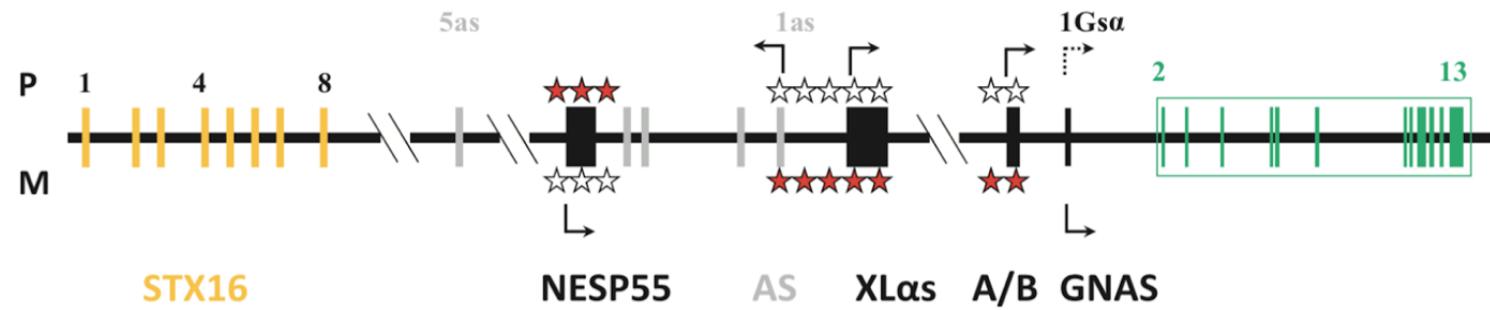
Secuenciación de Exoma

- Preparación de librerías utilizando el kit de enriquecimiento y captura “*SureSelect Target Enrichment System Human all exon V5 + UTRs*” de Agilent.
- Secuenciación en plataformas HiSeq 2000 de Illumina en dos empresas biotecnológicas (28 y 14 muestras respectivamente).



Deep-Seq

- 33 pacientes afectados por BWS, SRS, ANS, PHP y TS.
- Estudios previos: MS-MLPA.
- Diseño de sondas dirigido a *loci* de impronta.
- Preparación de librerías utilizando el método de captura “*SureSelect DNA*” de Agilent.
- Secuenciación en plataforma HiSeq 2000 de Illumina.
- Cobertura teórica de 1000x.



Pipelines bioinformáticas

Heriline

- Datos de exoma completo.
- Segregación familiar de variantes.
- Variantes en exones y zonas de splicing.

PeterBAM

- Datos de Deep-Seq.
- Variantes en exones, intrones y regiones intergénicas.
- Regiones repetitivas.
- CNVs.

Especificaciones técnicas

peterBAM

Scripts for probe-amplified deepseq analysis

Arguments to run peterBAM:

- Path to bed file
- Path to original fastq files
- Path to project

```
usage: utils.py [-h] [--test] [--debug] [--verbose] [--ori_fastq ORI_FASTQ]
                [--ori_bed ORI_BED] [--project_path PROJECT_PATH]
                [--show_steps]

Pipeline for DeepSeq analysis 'peterBAM'

python3.5 utils.py -fq fastq_path -b bed_file -p project_path

optional arguments:
  -h, --help            show this help message and exit
  --test, -t             Run pipeline on test mode
  --debug, -d            Run pipeline on debug mode
  --verbose, -v          Run pipeline with verbosity
  --ori_fastq ORI_FASTQ, -fq ORI_FASTQ
                        Set path to the original fastq directory for symlink
                        to project
  --ori_bed ORI_BED, -b ORI_BED
                        Set path to the original bed/manifest file for symlink
                        to project
  --project_path PROJECT_PATH, -p PROJECT_PATH
                        Set path to the project directory
  --show_steps, -s       Show steps of the pipeline
```

<https://github.com/carolinamonzo/>

- Heriline
- peterBAM
- impronta_scripts

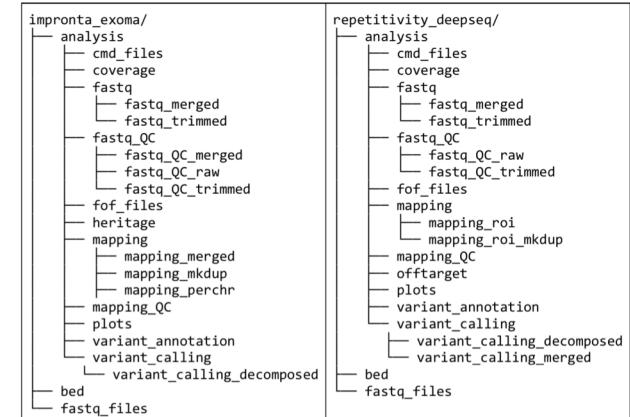
Ejecución de las pipelines

```
cmc@mochi:/nfs/production2d/cmc_projects_tmp/heriline
$ python3.5 heriline_wrapper.py -fq /media/qnapugdg8tb_2b/impronta_ori-2017-11-
27/EXOMA/rawdata/fastq -b ../impronta_exoma/bed/probes_exons_agilentv5Biomart.bed -p
..../impronta_exoma/
```

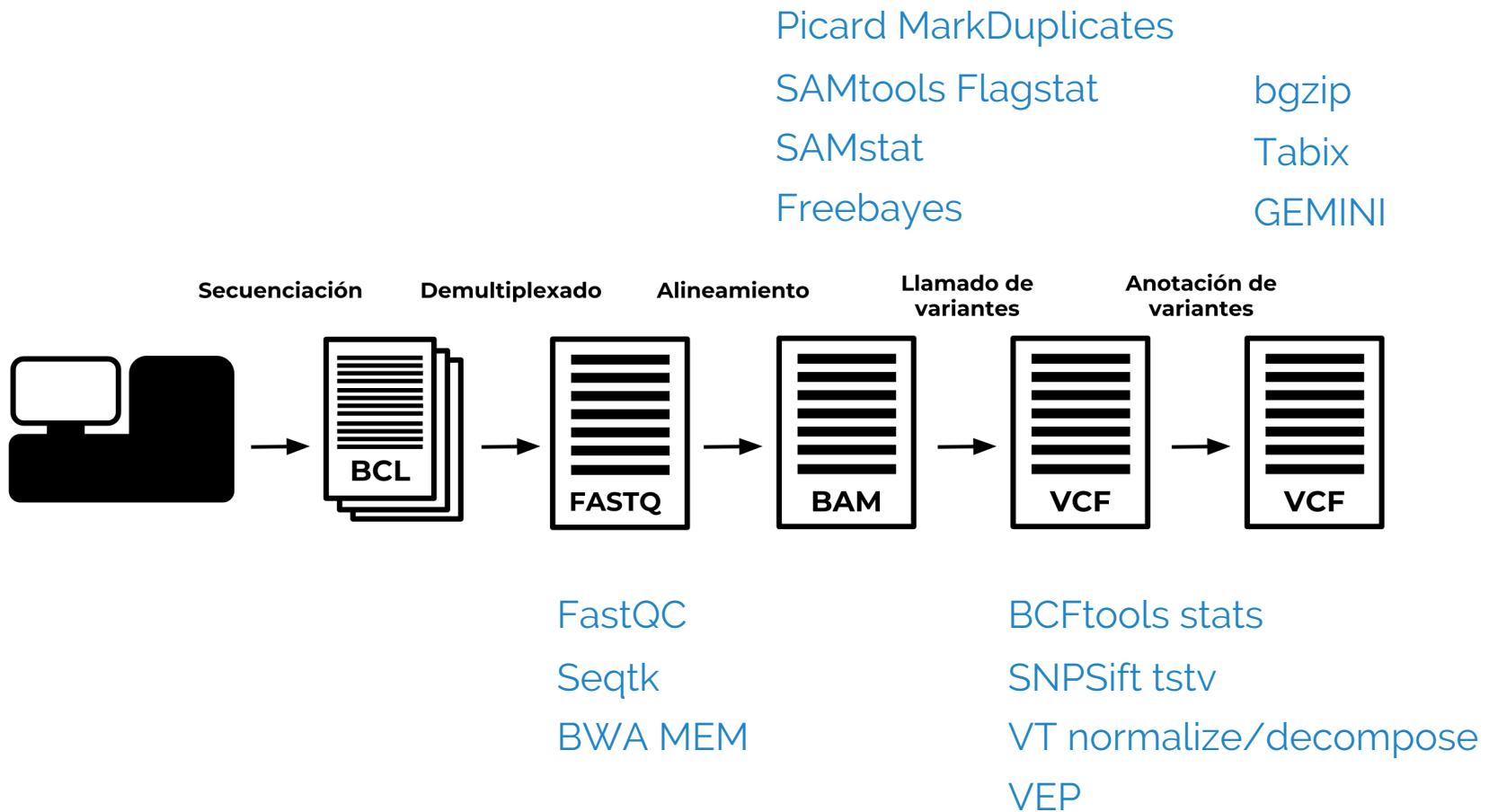
```
cmc@brownie:/nfs/production3b/cmc_projects_3b/peterBAM
$ python3.5 peterBAM_wrapper.py -fq /media/qnapugdg8tb_2b/impronta_ori-2017-11-
27/DEEPSEQ/ -b
/nfs/production3b/cmc_projects_3b/repetitivitiy_deepseq/bed/final_roi_woutRP.bed -p
..../repetitivitiy_deepseq/
```

```
cmc@mochi:/nfs/production2d/cmc_projects_tmp/heriline
$ python3.5 utils.py -fq /media/qnapugdg8tb_2b/impronta_ori-2017-11-
27/EXOMA/rawdata/fastq -b ../impronta_exoma/bed/probes_exons_agilentv5Biomart.bed -p
..../impronta_exoma/
```

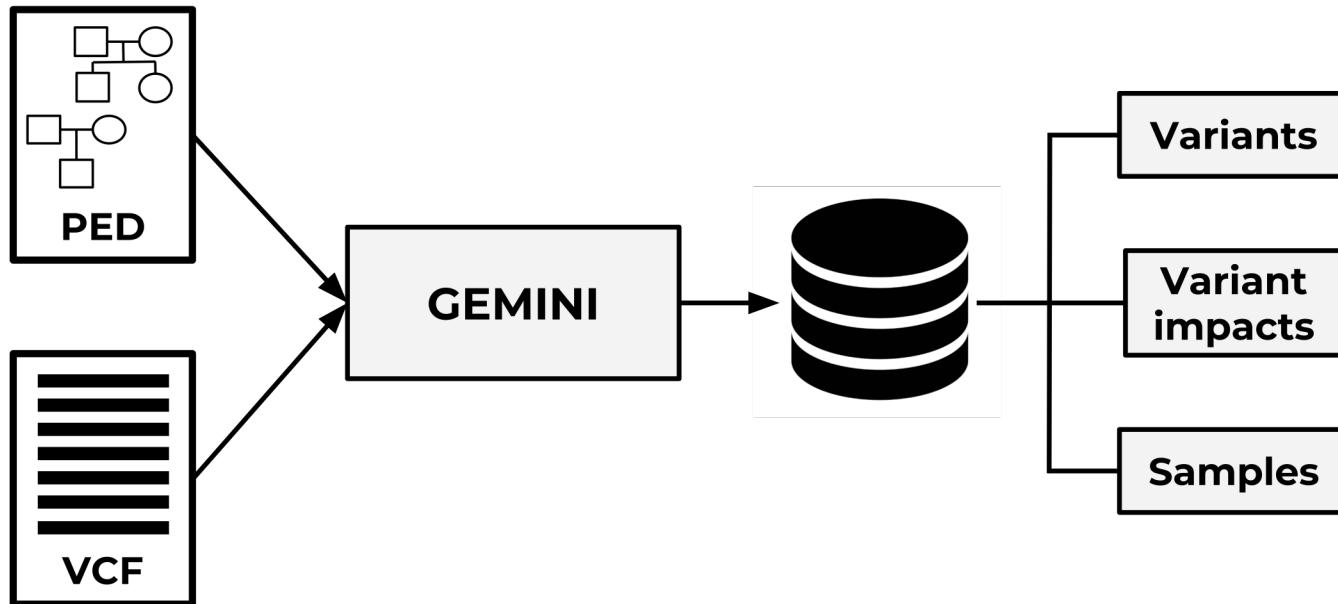
```
cmc@mochi:/nfs/production2d/cmc_projects_tmp/heriline
$ python3.5 mark_duplicates.py -p ..../impronta_exoma/
[INFO]: CMD_FILE -
./impronta_exoma/analysis/cmd_files/cmd_mark_duplicates_20180607_09-49-00.sh
[CMD]: parallel --joblog
./impronta_exoma/analysis/mapping/mapping_mkdup/mark_duplicates_20180607_09-49-
00.log -j 10 ::::
./impronta_exoma/analysis/cmd_files/cmd_mark_duplicates_20180607_09-49-00.sh
[INFO]: FOF_FILE -
..../impronta_exoma/analysis/fof_files/marked_duplicates_bam_20180607_19-40-24.fof
```



Flujo de trabajo general



Base de datos de variantes



Frecuencia poblacional europea < 0.1.

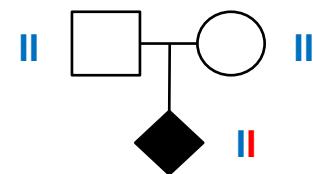
Valor de patogenicidad CADD > 20.

Impacto de las variantes de media a alta severidad.

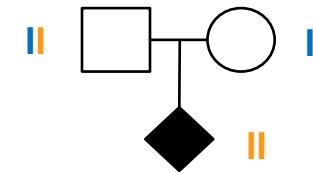
Elementos específicos de Heriline

- Priorización de variantes dirigida a patrones de herencia familiar.

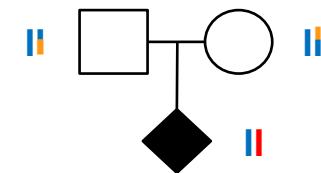
De-novo



Autosómico recesivo



Heterocigoto compuesto



Elementos específicos de PeterBAM

- Análisis de la mapabilidad y el *off-target* de las regiones de interés.
- Selección de lecturas de alta confianza.



- % GC.
- K-meros de 50 – 200.
- *Off-target*.
- Zonas repetitivas.
- Homología en *loci* de impronta.

Elementos específicos de PeterBAM

- Identificación de CNVs.



Validación de las *Pipelines*

Heriline

- Comparación de variantes “Gold Standard” del GIAB, individuo NA12878.

PeterBAM

- Secuenciación Sanger.
- MS-MLPA.

4. Resultados

Control de calidad

Exoma

Metrics	RAW FASTQ			TRIMMED FASTQ		
	PASS	WARN	FAIL	PASS	WARN	FAIL
Per base N content	73	3	6	80	0	2
Per base sequence content	24	2	56	26	0	56
Per base sequence quality	66	9	7	67	8	7
Sequence Duplication	45	35	2	45	37	0

Deep-Seq

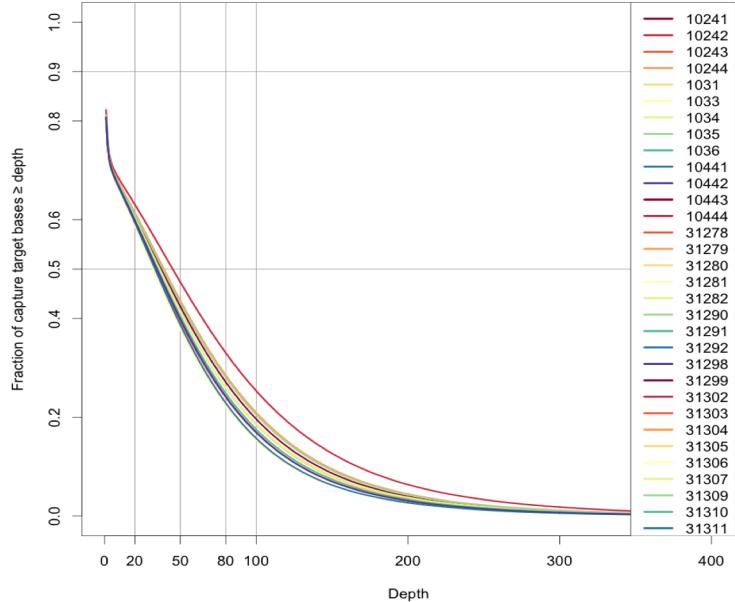
Metrics	RAW FASTQ			TRIMMED FASTQ		
	PASS	WARN	FAIL	PASS	WARN	FAIL
Per base N content	66	0	0	66	0	0
Per base sequence content	66	0	0	66	0	0
Per base sequence quality	62	4	0	66	0	0
Sequence Duplication	66	0	0	66	0	0

Control de calidad

Alineamiento

	Mean Total Reads	Mean % Reads Aligned	Mean % Reads Quality > 30
Exome	36,580,700	99.914	82.9
Deep-Seq-pre	36,043,120	99.764	27.4
Deep-Seq-post	7,278,219	100	99.3

Exome Coverage



Profundidad de cobertura

Exoma

46 - 105x

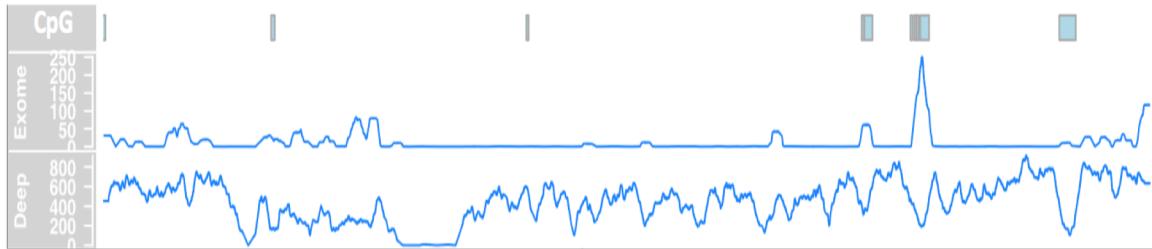
Deep-Seq

269x

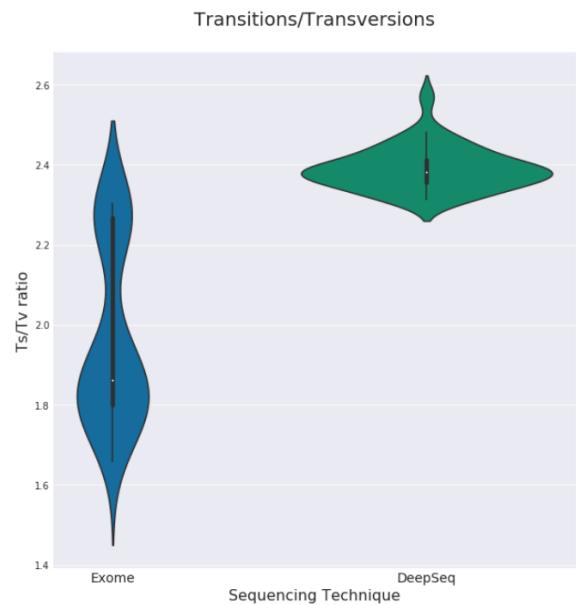
Control de calidad

Llamado de variantes

	SNV	Indel	Substitution	Deletion	Insertion
Exome	4,186,590	2,397	31,150	14,584	9,686
Deep-Seq	14,870	1,756	273	65	45



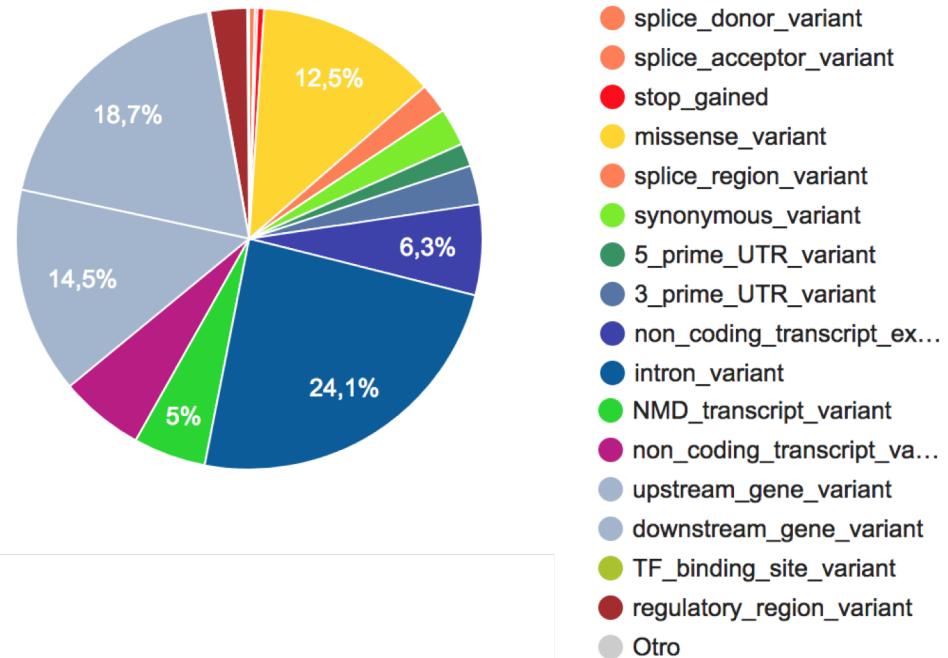
Locus GNAS



Análisis de variantes de Exoma

4.244.407 variantes únicas.

- 1 Paciente portador de mutación de cambio de sentido *de-novo* en E1-GNAS.
- 3 Pacientes portadores de mutación de cambio de sentido en E1-GNAS, herencia materna.
- 2 Pacientes portadores de mutación de cambio de sentido en E5-STX16, herencia materna.

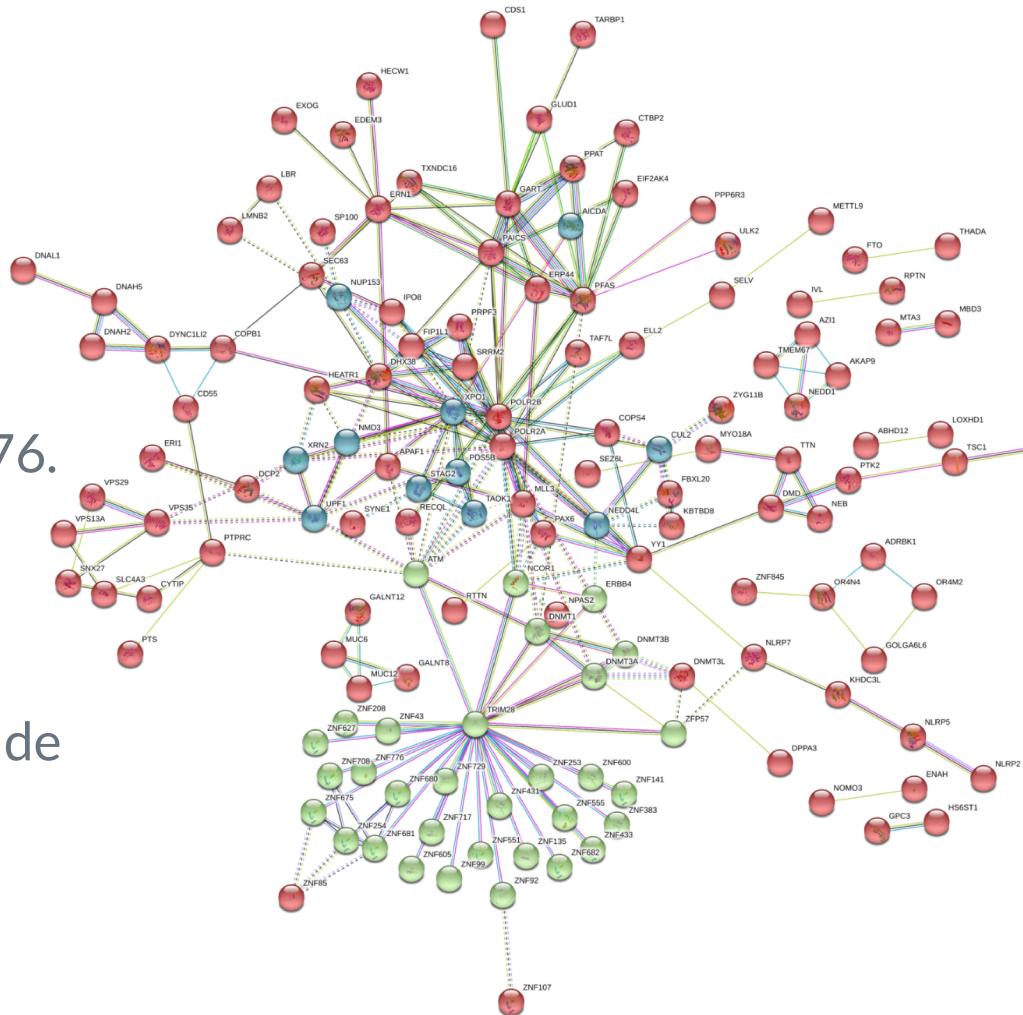


Análisis de variantes de Exoma

Impact	Compound Het	Recessive	De Novo
Splice acceptor variant	5	1	198
Splice donor variant	2	0	159
Start lost	0	0	33
Stop gained	17	4	510
Stop lost	0	0	4
Total	24	5	904

Análisis de variantes de Exoma

- 241 nodos/genes.
- 212 aristas/relaciones.
- Promedio de conexiones 1,76.
- Coeficiente de agrupación local 0,362.
- P-valor de enriquecimiento de interacciones $1,11e-07$.

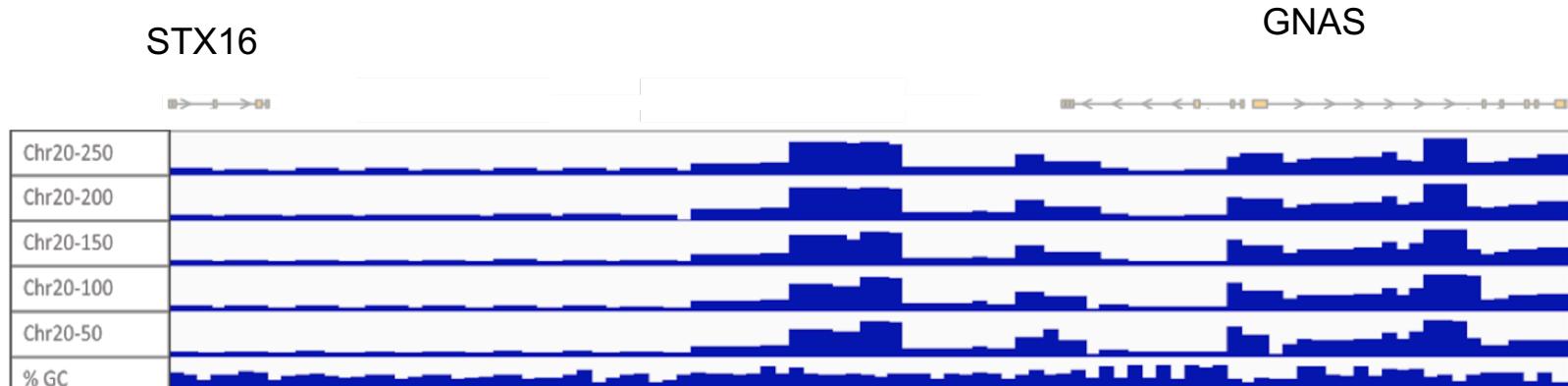


Validación de Heriline

	Total variants	SNPs	Indels	PPV	Sensibility
GIAB	3,392,227	2,967,659	437,792	-	-
Heriline	3,591,287	3,152,457	439,880	0.92	0.97

Valor Predictivo Positivo (PPV): considera "verdaderas positivas" a las variantes observadas tanto en los datos *gold standard* como en los resultados obtenidos en la pipeline, y "falsas positivas" las identificadas por la pipeline pero no incluídas en los datos *gold standard*.

Mapabilidad y off-target en Deep-Seq



Media del número de lecturas

Genoma completo

36.693.847

Off-target de las regiones de interés

24.504.423

Regiones de interés

11.937.437

Regiones repetitivas en Deep-Seq

- 836 sondas totales.
- 521 sondas solapantes con regiones repetitivas.

Media del número de lecturas

Lecturas repetitivas en *off-target* de la región de interés

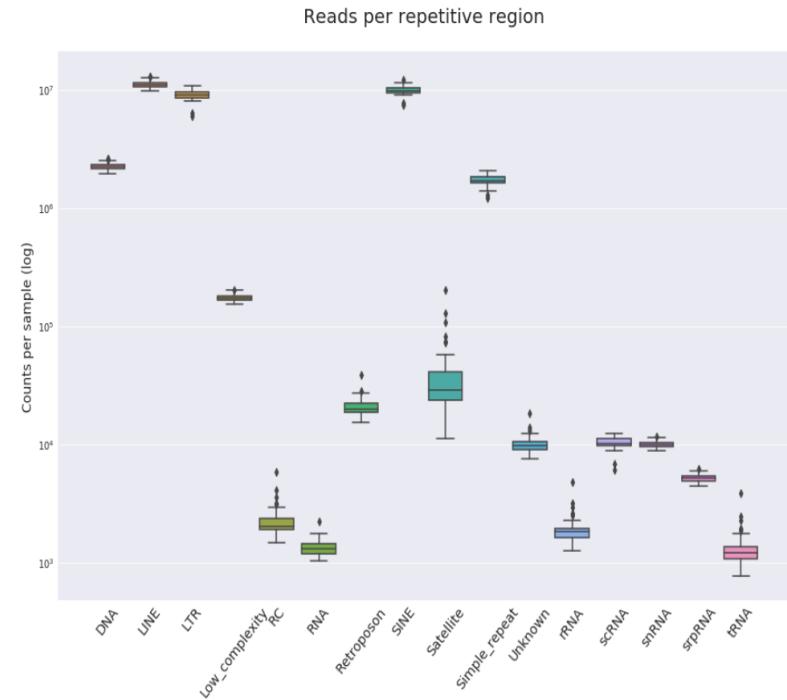
22.615.564

Lecturas repetitivas en la región de interés

6.173.926

Lecturas NO repetitivas en la región de interés

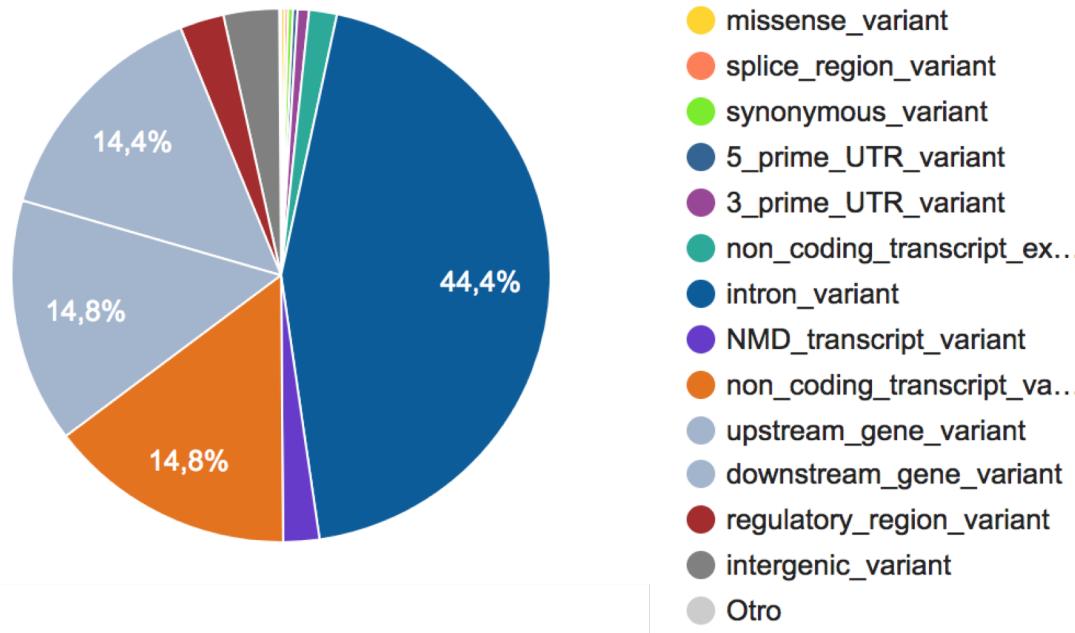
7.840.771



Análisis de variantes de Deep-Seq

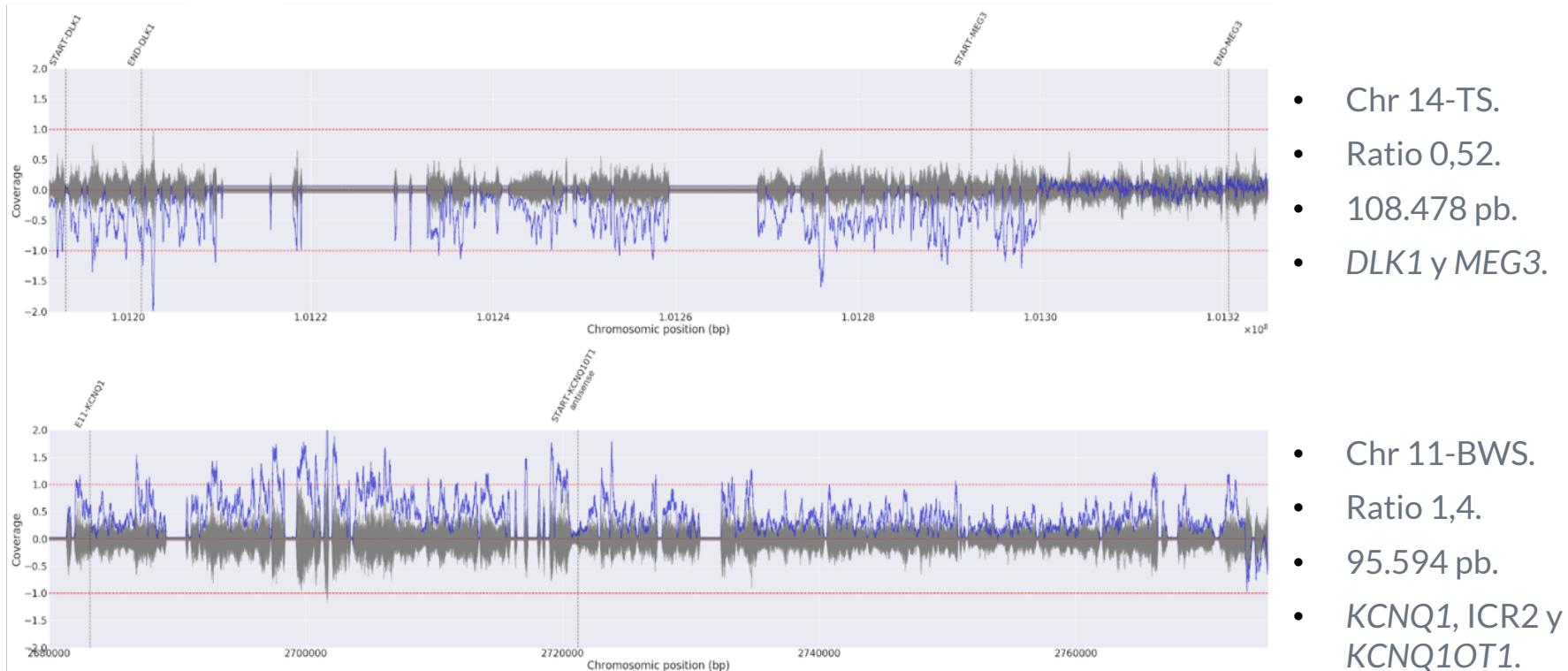
17.017 variantes únicas.

- 1 Paciente portador de mutación no descrita en E11-GNAS.
- 1 Paciente portador de mutación no descrita en la región de splicing de E11-GNAS.

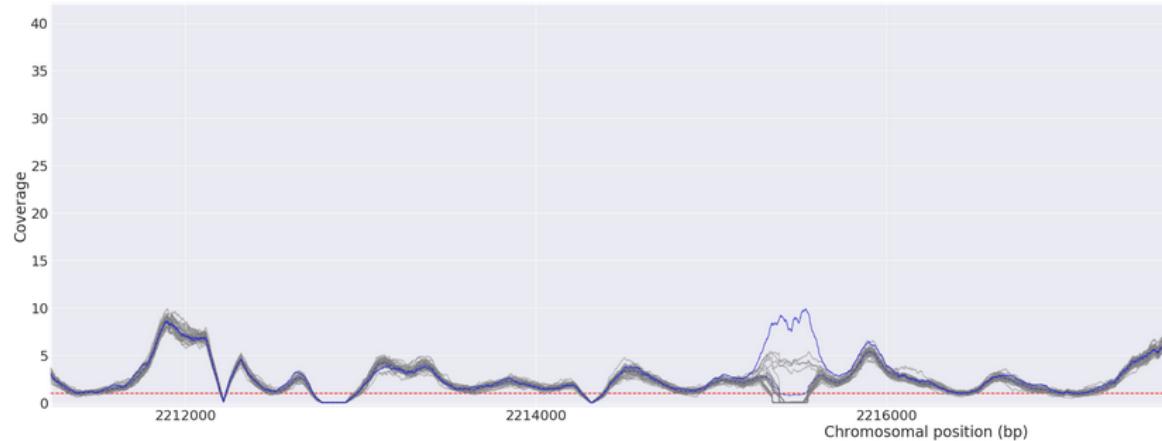


CNVs en Deep-Seq

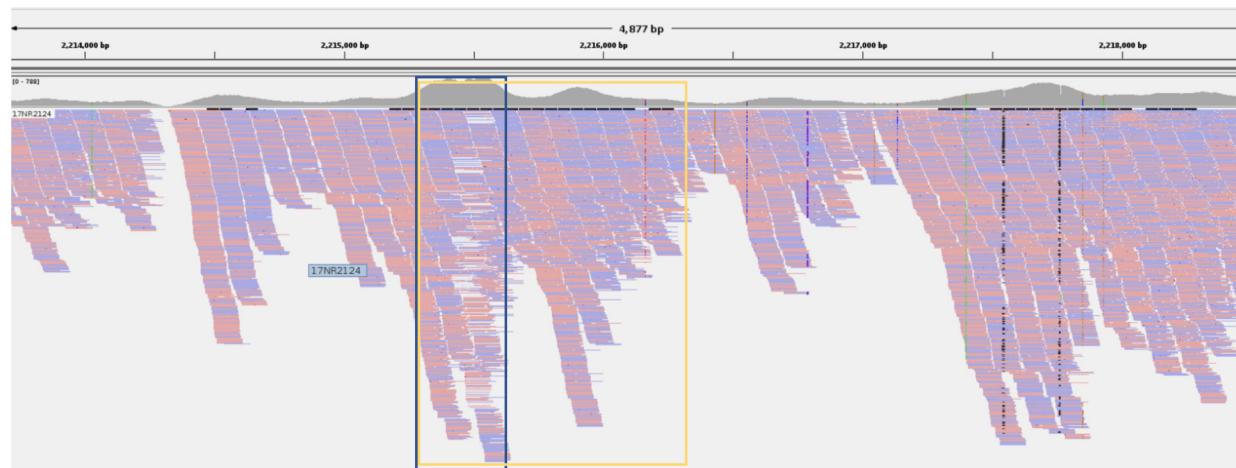
- 5 CNVs de interés.
- 7 Puntos de corte en SINE/LINE.



CNVs en Deep-Seq



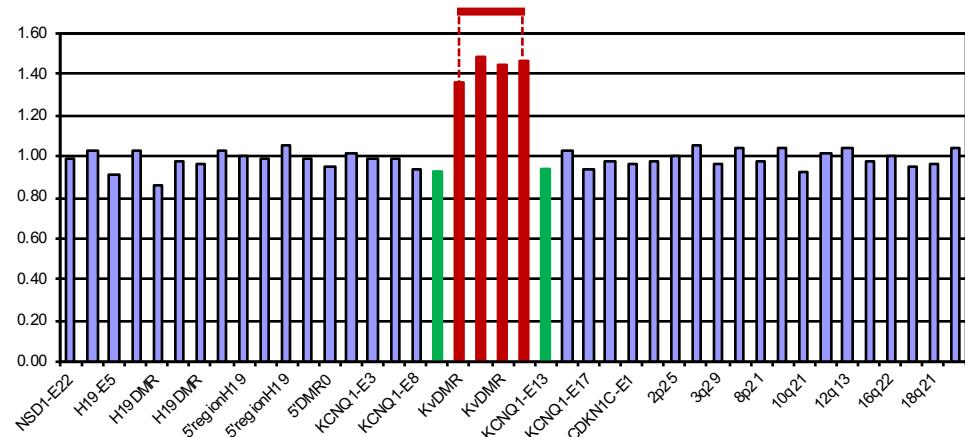
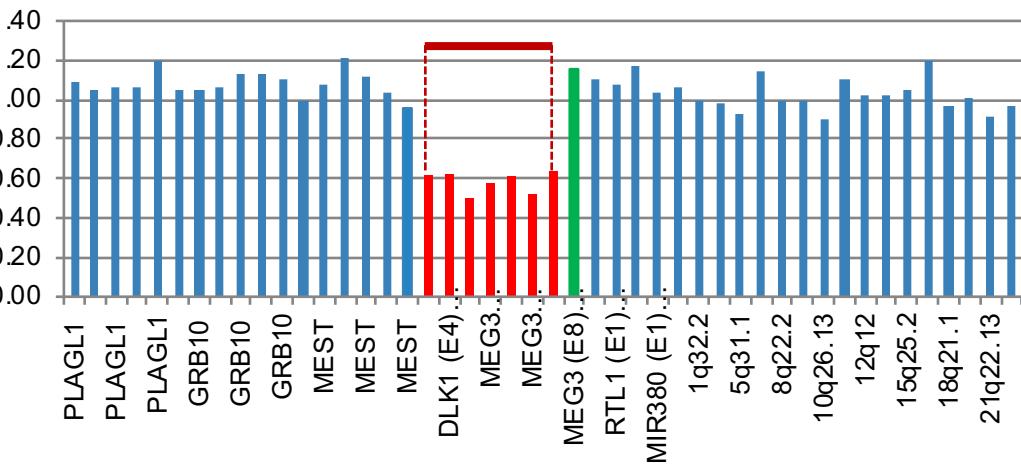
- Chr 11-BWS.
- Ratio 1,9.
- 455 pb.
- ICR1, CTCF.



Validación de CNVs y variantes en Deep-Seq

- Las variantes identificadas en GNAS, se comprobaron mediante Sanger.
- Las CNVs clasificadas como probablemente patogénicas, se comprobaron mediante MS-MLPA.

	Variantes	CNVs	Total
Pacientes	2	5	33



4. Conclusiones

Heriline

- Se ha desarrollado una *pipeline* bioinformática automatizada y reproducible, para realizar el estudio de datos de secuenciación de Exoma completo y analizar tanto las mutaciones detectadas como sus patrones de herencia.
- Este método bioinformático ha sido validado mediante la aplicación de un método basado en la comparación con un set de variantes “Gold Standard” para determinar la fiabilidad de la *pipeline*.
- El estudio de la interacción entre proteínas de genes de interés identificados como consecuencia de la ejecución de Heriline, ha permitido la identificación de genes candidatos que pueden estar actuando en ***trans*** en la regulación de la metilación.

PeterBAM

- Se ha puesto a punto una segunda *pipeline* bioinformática, PeterBAM, dirigida al estudio de datos de Deep-Seq. Esta metodología ha permitido reducir el ruido producido por secuencias repetitivas, y analizar variantes exónicas, intrónicas e intergénicas, además de detectar CNVs.
- Esta metodología ha sido validada mediante la comparación de los resultados obtenidos en los pacientes, con datos producidos con tecnologías ortogonales; MS-MLPA para las CNVs y secuenciación Sanger para las variantes.
- Mediante la ejecución de PeterBAM, se han identificado alteraciones genéticas (CNVs y mutaciones puntuales) en regiones en *cis* implicadas en la regulación de la impronta y el desarrollo del fenotipo de la enfermedad.
- La *pipeline* PeterBAM ha facilitado el establecimiento de los puntos de corte de las CNVs identificadas, proporcionando así información detallada de la amplitud de las regiones afectadas.

¡Gracias!

Carolina Monzó

Trabajo de Fin de Máster en Bioinformática

21-09-2018