

Reporte TumorSec: Control de calidad de los datos generados.

En el siguiente reporte se presentan diversas métricas de calidad correspondiente a la corrida 200523_TumorSec este informe fue generado el día 25/05/20 por el pipeline bioinformático de TumorSec desarrollado por el laboratorio de Genómica del Cáncer, Universidad de Chile. Estas métricas representan una guía para verificar la calidad de los resultados obtenidos por cada muestra en la secuenciación.

2.- Control de calidad de la corrida:

2.1.- Información de la corrida:

Nombre kit	MiSeq Standard Reagent Kit V2
Longitud de Lecturas	2x150 bp
Maximum Output	5.1 Gb
Cantidad de Lecturas	24-30 Millones
Calidad de Lecturas	80% > Q30.
Modo	Paired End
Equipo	MiSeq
N° Serie	M03158
Ciclos	2x150
Kit Preparación Librería	SeqCap EZ HyperCap Roche
Regiones Blanco	326,00
Largo Total Regiones Blanco (pb)	88.277,00

Tabla 1: Información del kit de secuenciación y regiones blanco de TumorSec

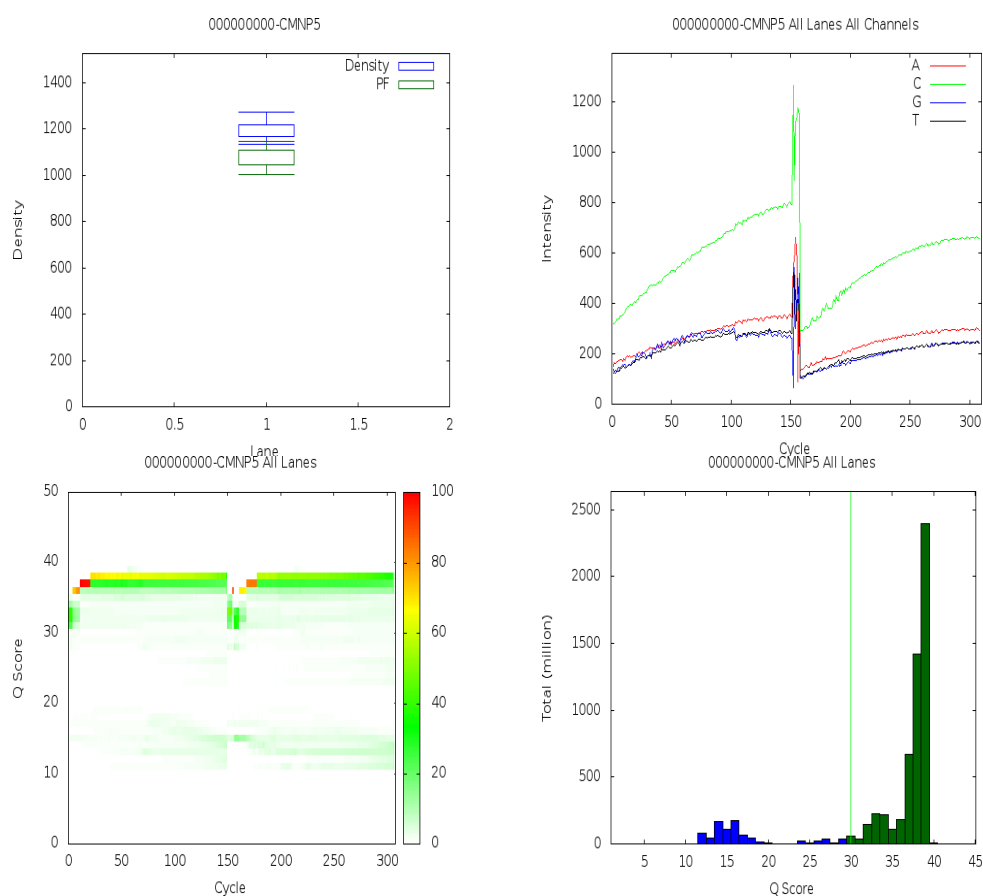


Imagen 1: Métricas de BaseSpace Illumina MiSeq. A: Densidad de clusters (K/mm2) por línea. B: Gráfico de intensidad por ciclo. C: Gráfico de score de calidad por ciclo. D: Diagrama de distribución de QScore, número de bases por score de calidad

2.1.- Métricas según Lecturas:

Read	Rendimiento (Gb)	Rend Esperado (Gb)	Intensidad	%>=Q30	Phas/Prephas (%)
Read1	3.07	3.07	149	90.3	0.095/0.076
Read2(l)	0.1	0.1	562	92.98	0.000/0.000
Read3	3.07	3.07	133	82.85	0.082/0.071
Non-indexed	6.15	6.15	141	86.57	nan
Total	6.25	6.25	281	86.68	nan

Tabla 2: Métricas de las lecturas de BaseSpace Illumina

2.2.- Métricas según Línea:

Línea	Densidad (K/mm2)	Clúster PF (%)
Línea 1	1197+/-39	89.87+/-1.24

3.- Control de calidad de los datos generados.

3.1.1- Tabla resumen de las métricas de calidad

Sample	R_before trimming	R_after trimming	% of trim	% of dups	Reads On-Target	% On-Target	Mean read coverage	% uniformity
Undetermined_S0	1396546.0	902484.0	35.38	51.61	168397.00	37.59	119.88	97.24
1590TFFPE_S1	5625220.0	4480184.0	20.36	65.36	317106.00	19.89	206.53	83.44
1692FFPE_S2	7391414.0	5699186.0	22.89	64.61	446966.00	21.52	272.13	80.37
1765TFFPE_S3	9158726.0	7905712.0	13.68	67.40	488364.00	18.54	337.76	82.52
2049TFFPE_S4	8309236.0	6884706.0	17.14	68.26	443216.00	19.83	296.78	90.49
1665FFPE_S5	6742150.0	6344962.0	5.89	68.77	368464.00	18.31	249.11	96.93
FA6060FFPE_S6	2370292.0	2025898.0	14.53	40.80	493905.00	40.08	249.51	97.55

Tabla 3: Métricas de calidad. Sample: Muestras. R_before trimming: Número de lecturas iniciales. R_after trimming: Número de lecturas después de la poda. % of trim : Porcentaje de lecturas eliminadas. % of dups: Porcentaje de lecturas duplicadas. Reads On-Target: Número de lecturas en regiones blanco. % On-Target: Porcentaje en regiones blanco. Mean read coverage: Profundidad promedio de lecturas en regiones blanco. % Uniformity: Porcentaje de lecturas en regiones blanco, donde su profundidad promedio está entre el rango $[2 \times \text{mediana}, \text{mediana}/2]$.

3.1.2- Profundidad y amplitud de las coberturas en las regiones blanco:

SAMPLE	MEAN_COVERAGE	MAX_COVERAGE	PCT_100X	PCT_200X	PCT_300X	PCT_400X	PCT_500X
Undetermined_S0	112.29	333.0	59.77	1.19	0.01	0.00	0.00
1590TFFPE_S1	184.42	503.0	81.20	42.21	9.62	0.50	0.01
1692FFPE_S2	237.13	619.0	82.51	55.21	33.81	13.68	1.55
1765TFFPE_S3	307.21	747.0	99.77	67.53	48.30	30.50	8.14
2049TFFPE_S4	268.68	664.0	99.38	65.42	39.89	14.96	1.30
1665FFPE_S5	232.82	468.0	98.10	69.46	14.28	0.50	0.00
FA6060FFPE_S6	224.19	992.0	96.91	60.03	14.78	2.27	0.70

Tabla 4: Métricas de cobertura en regiones targets

3.1.3- Amplitud de regiones blanco con cobertura mínima de 300X

SAMPLE	Prom_pct_amp	Nº < 70 %	Nº < 80 %	Nº < 90 %	Nº < 100 %
Undetermined_S0	0.20	326	326	326	326
1590TFFPE_S1	14.40	295	303	315	322
1692FFPE_S2	45.06	194	205	212	237
1765TFFPE_S3	56.81	146	152	162	173
2049TFFPE_S4	50.51	171	182	194	210
1665FFPE_S5	19.31	282	289	302	311
FA6060FFPE_S6	21.59	282	295	306	317

Tabla 5: Número de regiones blanco que poseen menos del 80% una profundidad mínima de 300X. Estas métricas también se observan para un 70%, 90% y 100% a un mínimo de 300X. Para resultados de mejor calidad, se espera un bajo número de regiones a 80% a un mínimo de 300X. El detalle de las regiones blanco se encuentra en los archivos excel y reporte complementario, generados por el pipeline TumorSec

3.2.- Calidad de bases por posición en fastq R1 y R2 sin filtrar: FASTQC

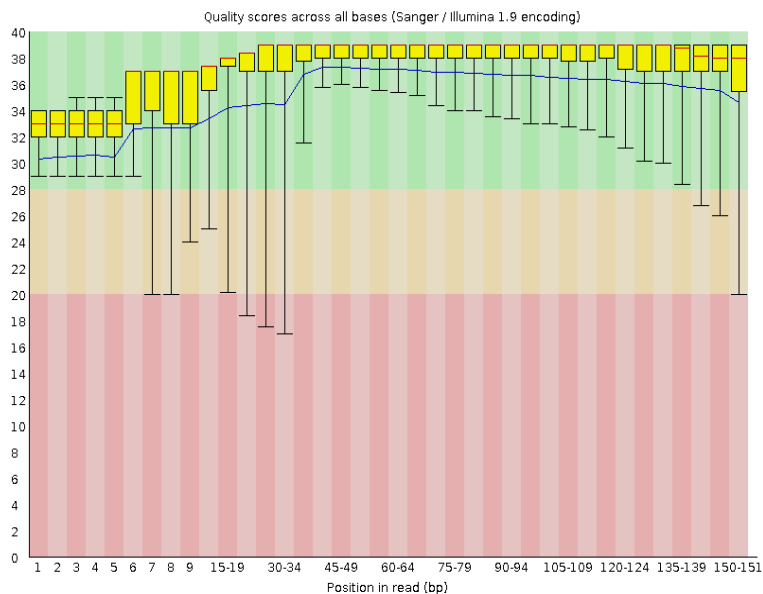


Imagen 2: Descripción general de los valores de calidad para todos los fastq R1. Se observa un gráfico de cajas con la distribución de las calidades por posición en la lectura. Este gráfico permite observar si los datos tienen algún problema antes de realizar cualquier análisis posterior.

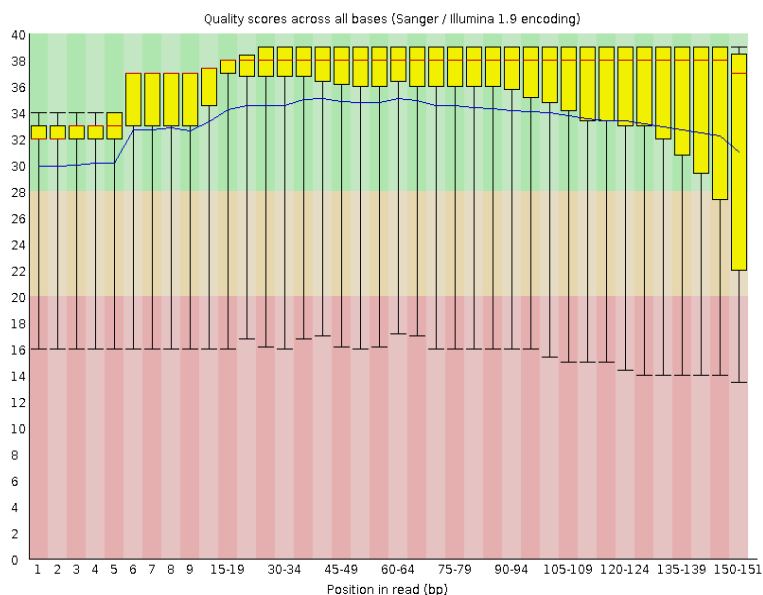


Imagen 3: Descripción general del rango de valores de calidad para todos los fastq R2 en cada posición.

3.3.- Cantidad de lecturas antes y después del trimming:

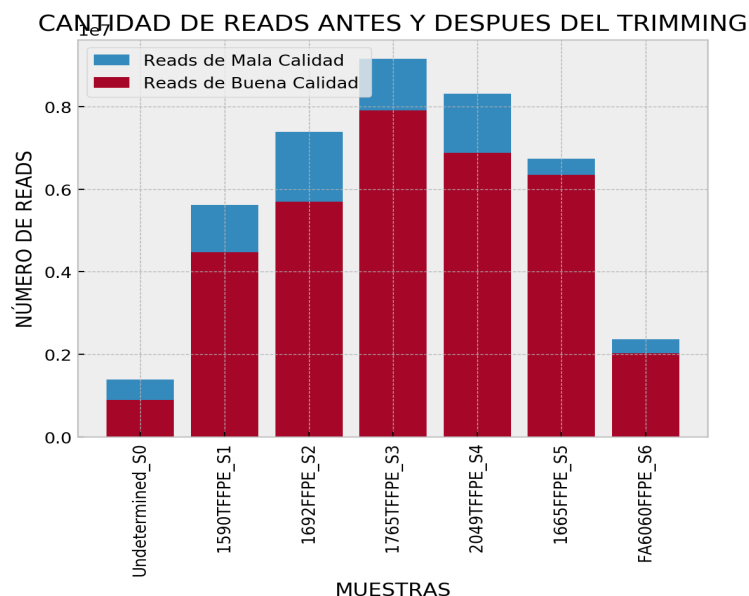


Imagen 4: Representación gráfica del trimming de datos realizado por fastp

3.4.- Calidad de lecturas antes y después del trimming:

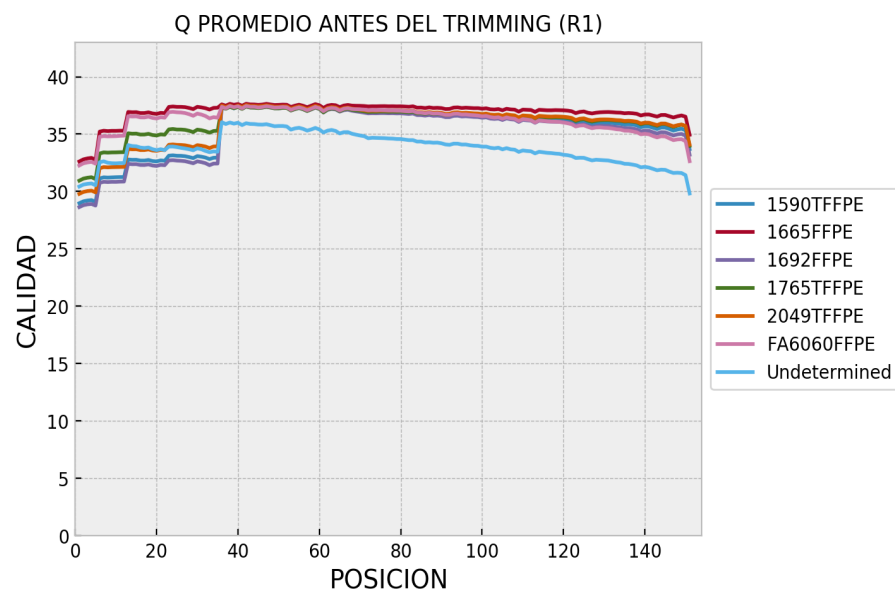


Imagen 5: Calidad promedio de lecturas por posición antes del trimming (R1)

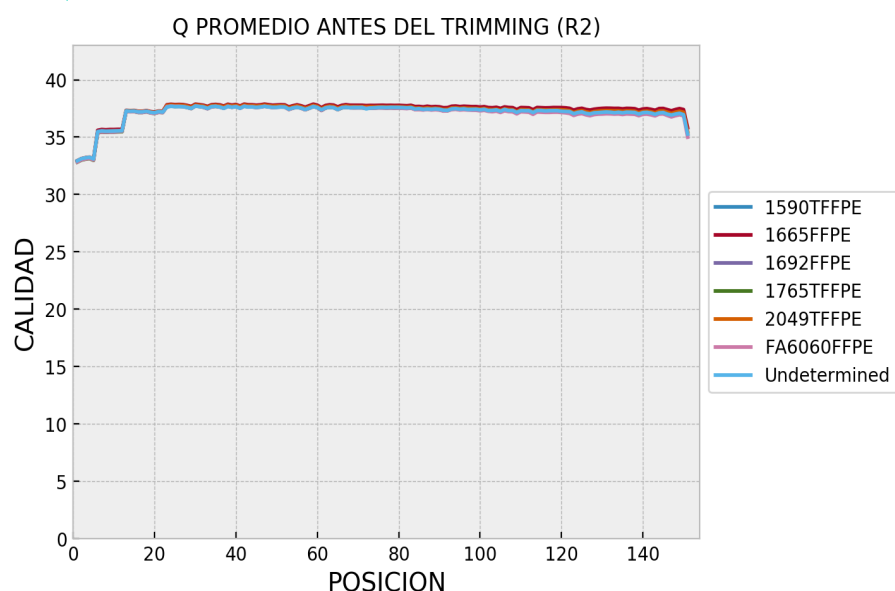


Imagen 6: Calidad promedio de lecturas por posición antes del trimming (R2)

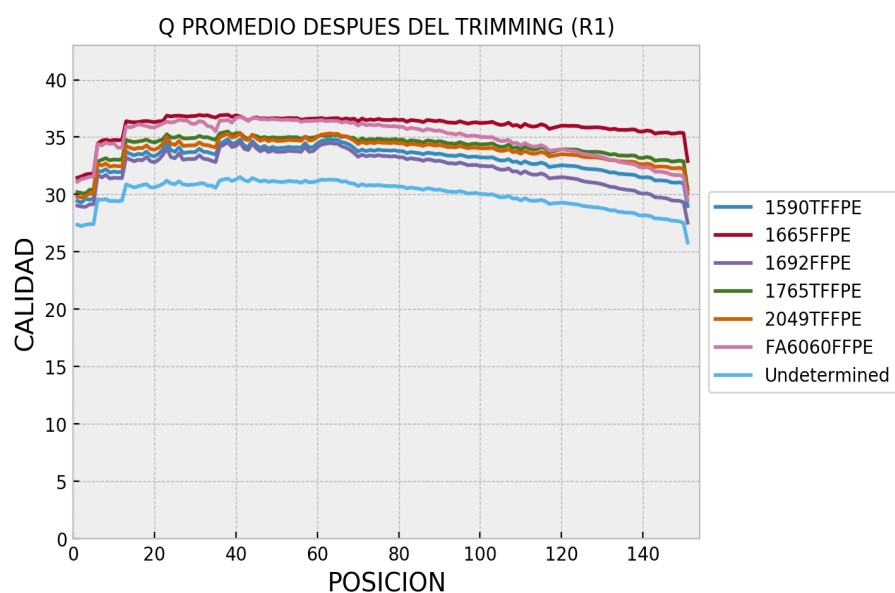


Imagen 7: Calidad promedio de lecturas por posición después del trimming (R1)

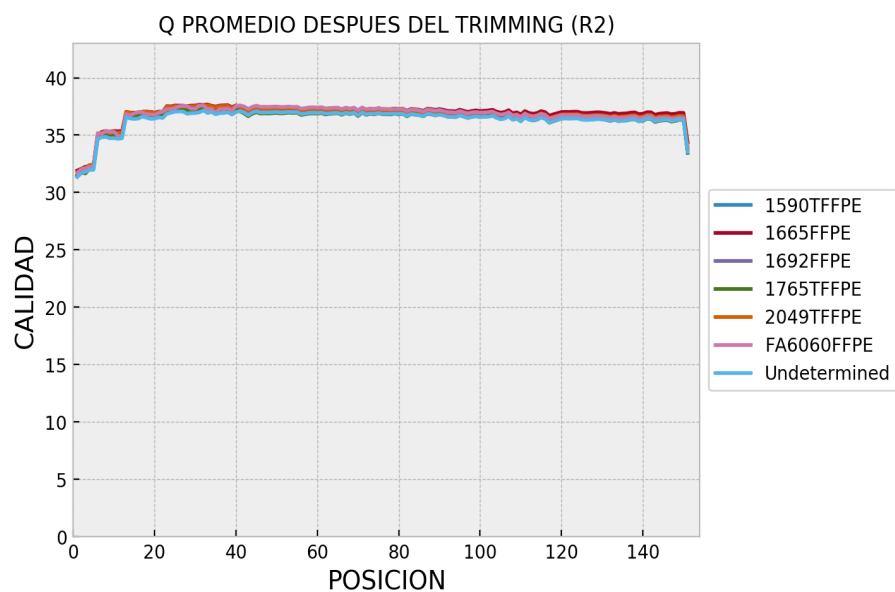


Imagen 8: Calidad promedio de lecturas por posición después del trimming (R2)

3.5.- Porcentaje de duplicados y tamaño del inserto

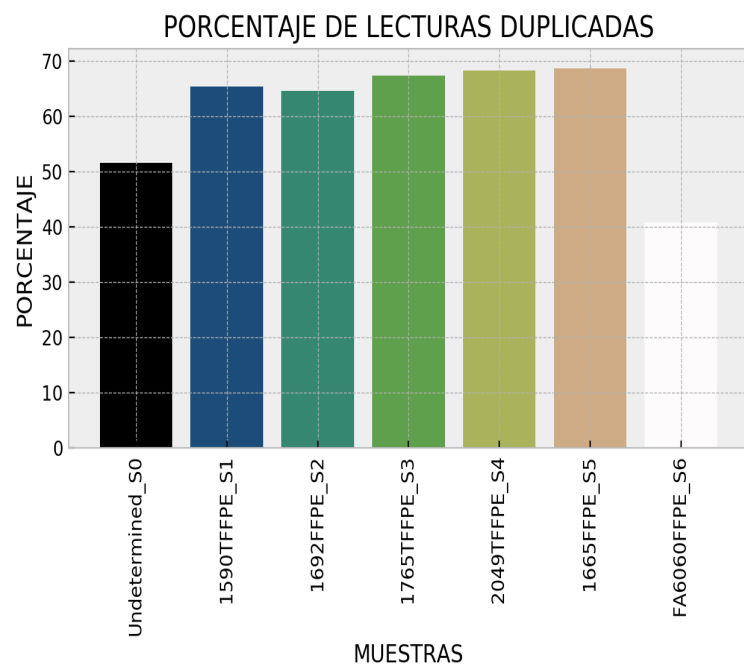


Imagen 9: Porcentaje de lecturas duplicadas, estas son marcadas e ignoradas en el llamado de variantes

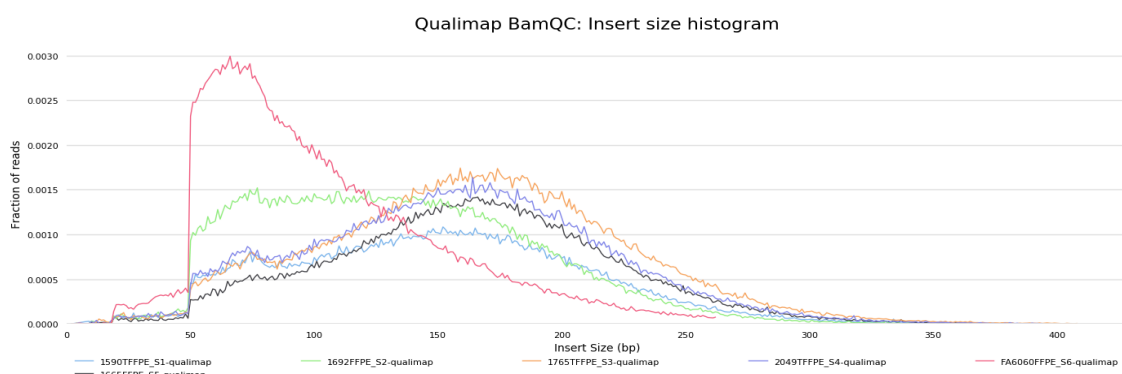


Imagen 10: Histograma de distribución del tamaño del inserto. Métricas obtenidas por qualimap

3.6.- Porcentaje de lecturas on-target y profundidad:

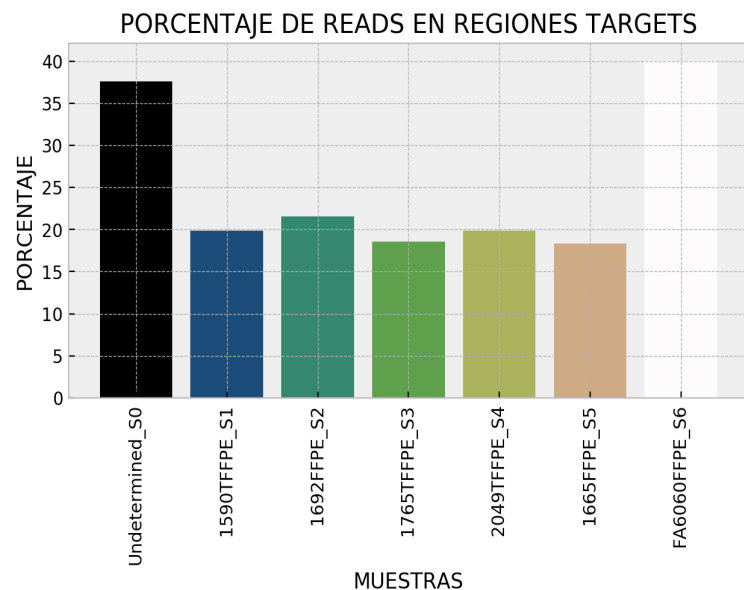


Imagen 11: Porcentaje de lecturas que se encuentran en regiones blancas por muestra

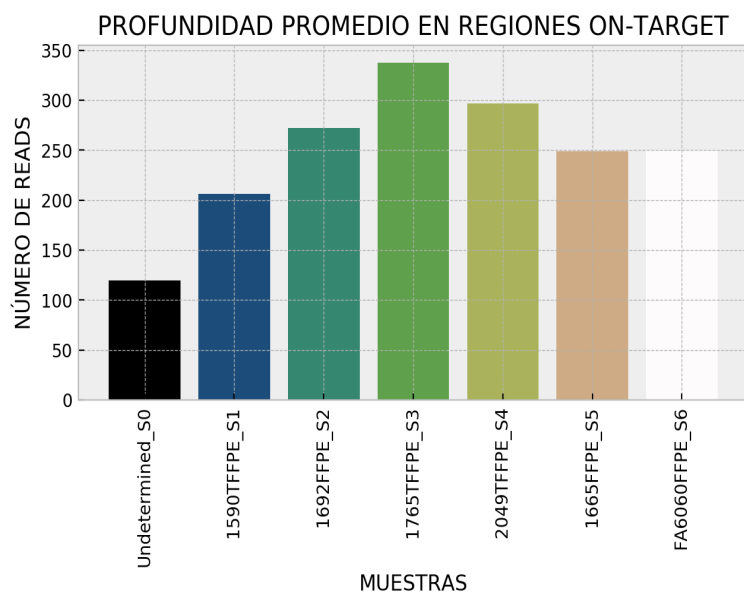


Imagen 12: Promedio de profundidad por muestras en regiones targets

3.6.- Uniformidad, intervalos con profundidad definida, bases con profundidad definida:

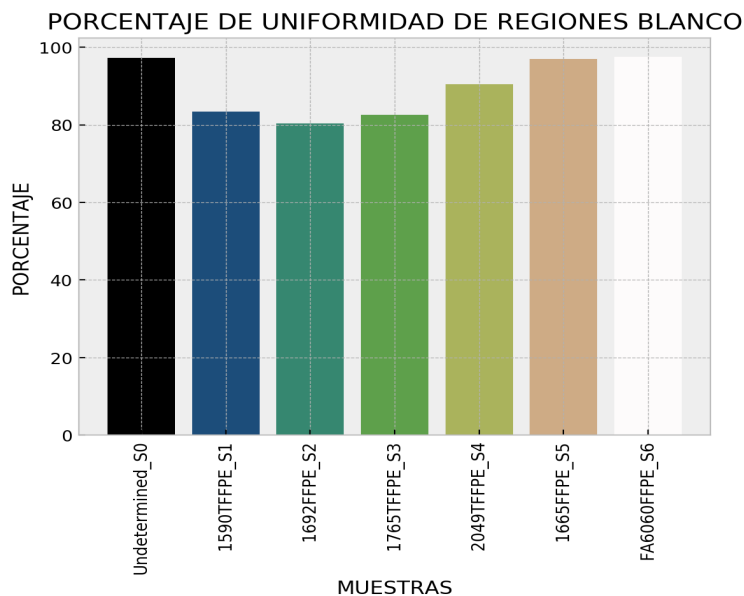


Imagen 13: Uniformidad de cobertura promedio de la corrida

CANTIDAD DE REGIONES BLANCO CON COBERTURA PROMEDIO POR UMBRAL

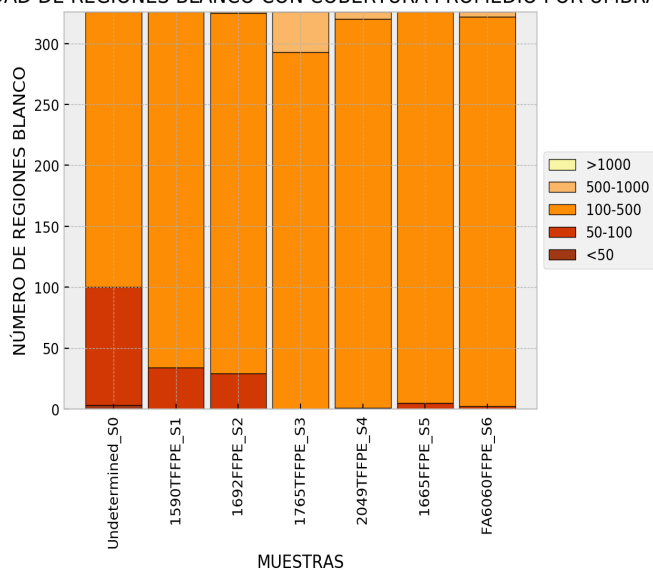


Imagen 14: Cantidad de coberturas promedio por rango para cada muestra

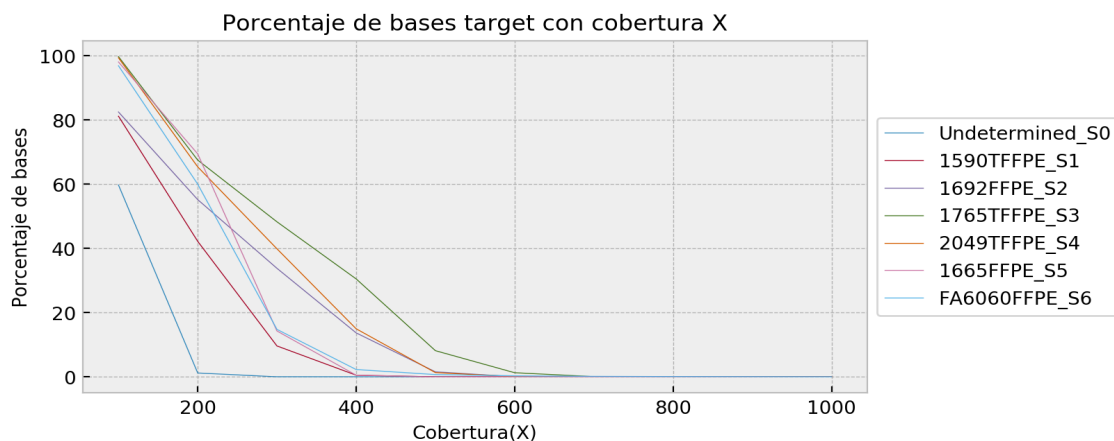


Imagen 15: porcentaje de bases con cobertura en un rango de 100X-1000X

4. Cobertura de bases por cada gen target

Cobertura de las regiones blanco por gen. Se agruparon las regiones blancos por gen target y se calcularon los porcentajes de bases que poseen una cobertura mayor o igual a 100X, 300X, 400X y 500X. Esta información, además se encuentra por muestra en el reporte suplementario del pipeline TumorSec.

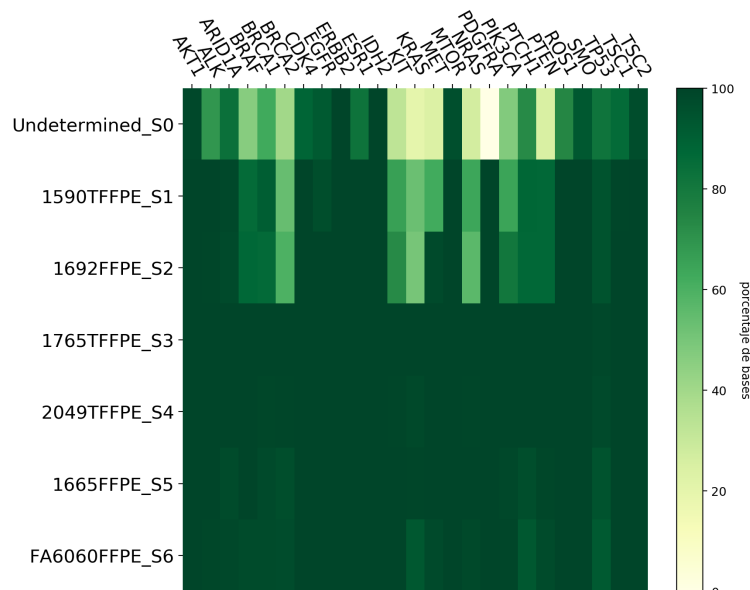


Imagen 16: Porcentaje de bases targets que poseen una cobertura mayor o igual a 100X

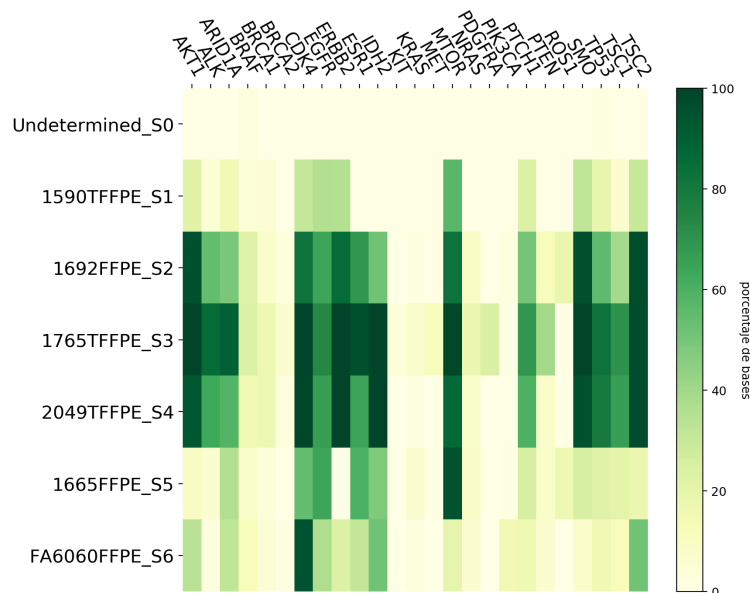


Imagen 17: Porcentaje de bases targets que poseen una cobertura mayor o igual a 300X

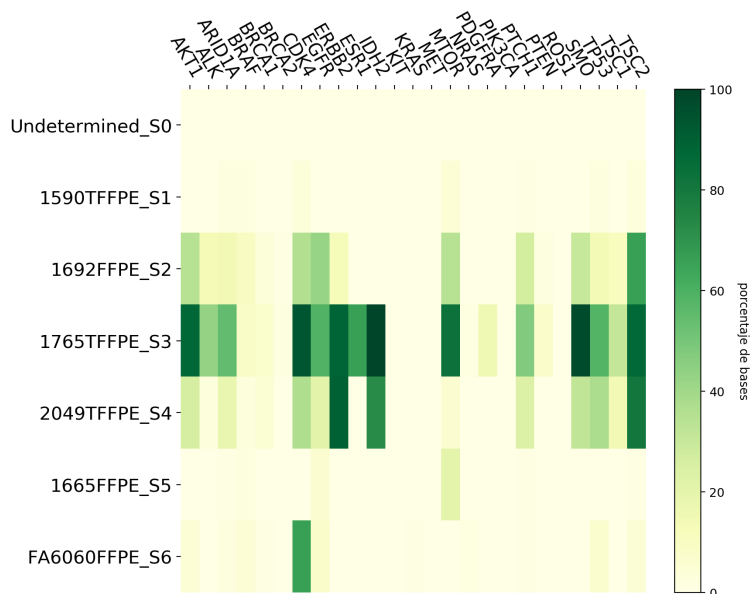


Imagen 18: Porcentaje de bases targets que poseen una cobertura mayor o igual a 400X

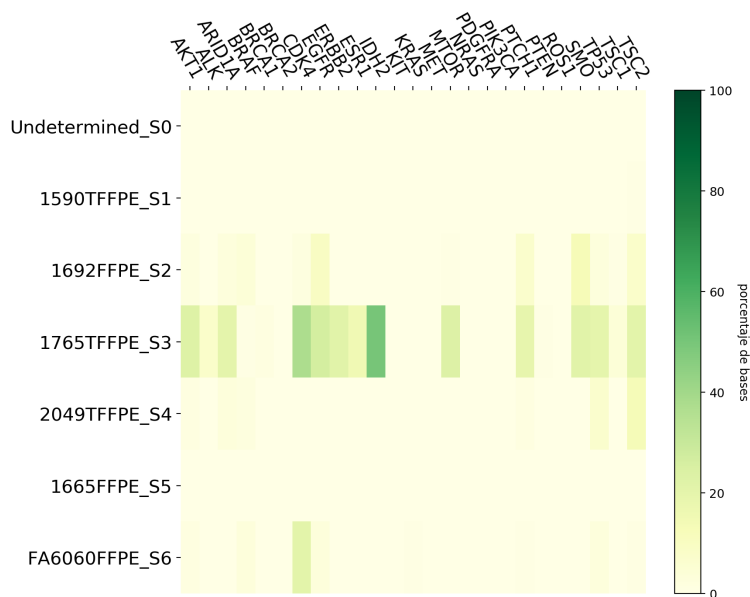


Imagen 18: Porcentaje de bases targets que poseen una cobertura mayor o igual a 500X