PREDICTION-ADR

University of Dundee

15th & 16th September, 2015

WP3 Targeted Exome Plus Sequencing

Cyrielle Maroteau, PhD

The PREDICTION-ADR project has received funding from the European Cummunity's Seventh Framework Programme (FP7/2007-2013) UNDER Grant Agreement no. 602108

Harmonisation and analyse of the sequences

Liverpool

Hiseq 2500

Uppsalla

Hiseq 2500

Dundee

Nextseq 500

Library preparation with the same exome target

Myopathy 250 cases 250 controls + **Angioedema** 250 cases and 250 controls

8 super-controls (statin & ACE) → Harmonisation

A high-level overview of NGS data processing:
From sequencing to variant file

Base calling
De-multiplexing

bcl2fastq v2.10.0.6

Fastq file : Raw NGS reads

Mapping Pipeline
-BWA-MEM
-BWA+Stampy

Alignment
or
Assembly

.sam file : Uncompressed text file
.bam file : Compressed & indexed file

SAM/BAM file :
Aligned NGS reads

Calling pipeline
-GATK Haplotypecaller
-Platypus

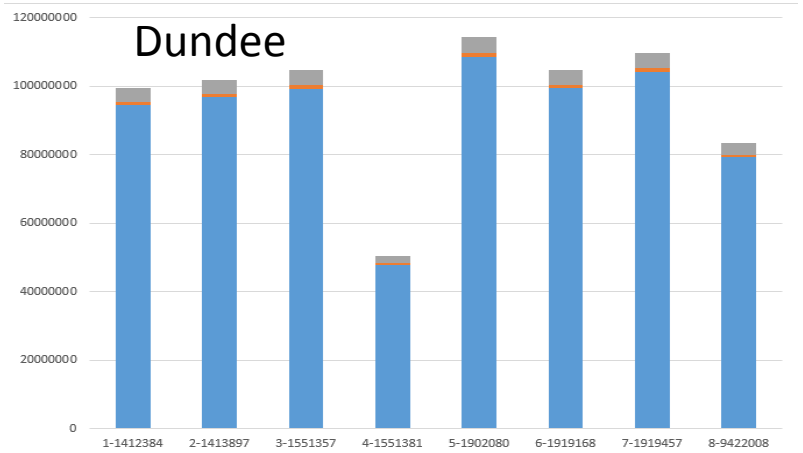
Variant calling

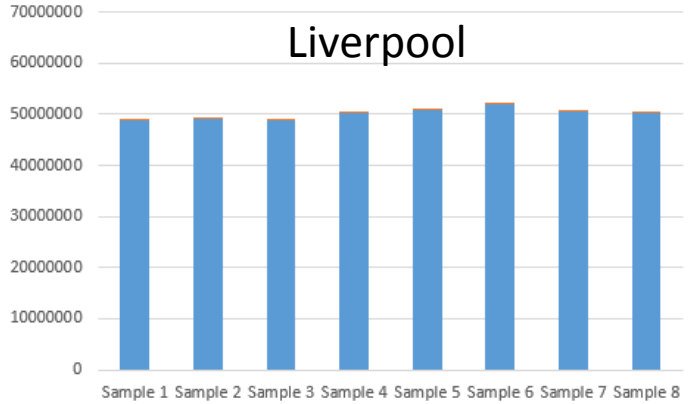
VCF files can be block-compressed & indexed

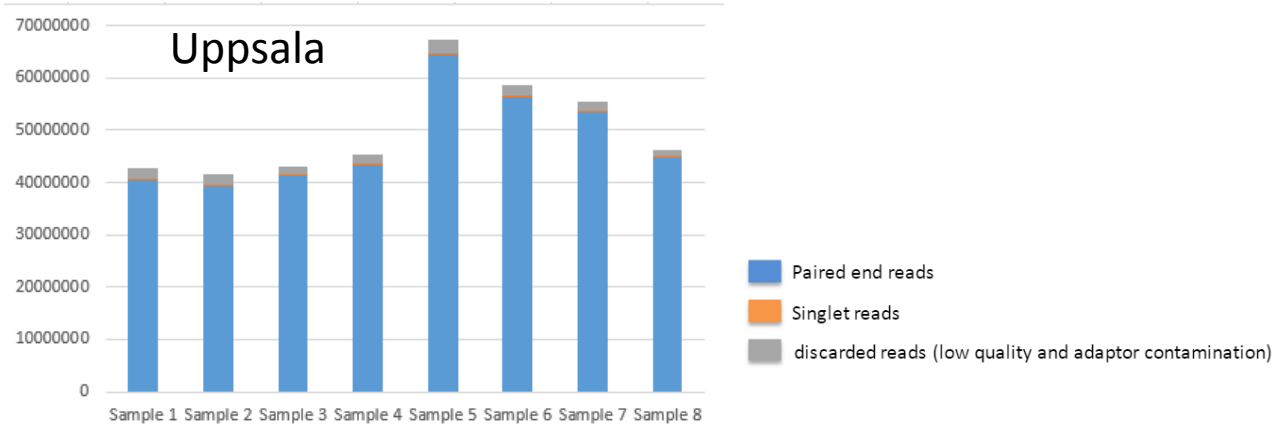
VCF file : Genomic variations

Comparable high quality reads obtained on the 3 different platforms

Number of reads in fastq files







Trimming of reads

1.Cutadapt version 1.2.1 (Martin, 2011): option “-O3”

2.Sickle version 1.200 : option qscore >20

3.Removing of 10bp reads

Comparison of Dundee 2 pipelines

- Dundee 8 SC sequences

  DP1: Dundee pipeline 1 (Basespace) BWA-MEM / GATK Haplotypecaller
  DP2: Dundee pipeline 2 (Cluster) Bwa-Stampy / Platypus

File Name	Multiplex per lane	% mapped		% at 20x coverage		% on target reads		% duplicates		No. of SNPs		No. of INDELs		Total no. of variants	
		DP1	DP2	DP1	DP2	DP1	DP2	DP1	DP2	DP1	DP2	DP1	DP2	DP1	DP2
1-ID9422008	8	98.50	98.94	96.20	92.20	86.10	77.25	20.00	14.37	40,886	36,986	2,925	3,816	43,811	40,802
2-ID1412384	8	98.30	98.80	95.20	92.00	85.80	79.66	20.70	18.60	41,102	36,682	2,962	3,540	44,064	40,222
3-ID1919457	8	98.60	98.87	95.40	94.00	83.20	79.00	15.30	17.77	40,710	37,053	2,931	3,810	43,641	40,863
4-ID1902080	8	98.30	98.93	96.80	93.90	85.00	80.35	18.80	18.71	40,907	36,797	3,054	3,345	43,961	40,142
5-ID1919168	8	98.40	98.88	95.80	93.10	85.40	79.98	18.70	18.85	40,447	36,540	2,942	3,316	43,389	39,856
6-ID1551381	8	98.60	98.85	96.80	87.80	86.30	79.98	20.00	10.55	41,283	36,102	3,040	3,393	44,323	39,495
7-ID1413897	8	98.50	98.86	91.60	92.90	85.00	79.20	12.00	16.88	39,946	36,714	2,694	3,807	42,640	40,521
8-ID1551357	8	98.30	98.74	96.30	93.40	84.10	77.50	18.90	17.01	40,400	36,216	2,920	3,797	43,320	40,013
AVERAGE	**8**	**98.44**	**98.86**	**95.51**	**92.41**	**85.11**	**79.12**	**18.05**	**16.59**	**40,710**	**36,636**	**2,934**	**3,603**	**43,644**	**40,239**

Sequencing centre	Starting DNA	Library protocol	PCR cycles	Sequencing Machine	Illumina chemistry	Reads
UoD	25ng	SureSelect Human all exon V5	8	Illumina Nextseq500	V1	2 × 150bp paired end reads

99% concordancy between the 3 labs

File Name	Multiplex per lane	% mapped			% at 20x coverage			% on target reads			% duplicates			No. of SNPs			No. of INDELs			Total no. of variants		
		UOD	UOL	UOU	UOD	UOL	UOU	UOD	UOL	UOU	UOD	UOL	UOU	UOD	UOL	UOU	UOD	UOL	UOU	UOD	UOL	UOU
1-ID9422008	8	98.94	99.83	98.92	92.20	90.60	86.20	77.25	89.07	68.05	14.37	6.40	5.47	36,986	36,344	35,812	3,816	3,289	3,145	40,802	39,633	38,957
2-ID1412384	8	98.80	99.86	99.30	92.00	90.40	82.00	79.66	80.28	63.07	18.60	7.00	5.39	36,682	36,395	35,683	3,540	3,307	3,146	40,222	39,702	38,829
3-ID1919457	8	98.87	99.88	99.08	94.00	90.60	88.90	79.00	89.22	64.16	17.77	6.78	5.39	37,053	36,278	35,947	3,810	3,321	3,120	40,863	39,599	39,067
4-ID1902080	8	98.93	99.84	98.84	93.90	89.90	90.80	80.35	82.10	60.04	18.71	5.91	5.53	36,797	36,429	36,193	3,345	3,728	3,154	40,142	40,157	39,347
5-ID1919168	8	98.88	99.85	98.94	93.10	90.80	88.80	79.98	88.79	61.74	18.85	6.32	5.49	36,540	36,278	35,973	3,316	3,297	3,089	39,856	39,575	39,062
6-ID1551381	8	98.85	99.89	99.58	87.80	90.60	84.40	79.98	88.90	63.86	10.55	7.17	5.48	36,102	36,338	35,701	3,393	3,159	3,037	39,495	39,497	38,738
7-ID1413897	8	98.86	99.86	99.54	92.90	90.40	82.20	79.20	88.76	65.06	16.88	6.94	5.52	36,714	36,100	35,070	3,807	3,259	3,067	40,521	39,359	38,137
8-ID1551357	8	98.74	99.87	99.54	93.40	89.90	84.60	77.50	86.97	68.84	17.01	6.47	5.48	36,216	35,577	34,936	3,797	3,314	2,998	40,013	38,891	37,934
AVERAGE	**8**	**98.86**	**99.86**	**99.22**	**92.41**	**90.40**	**85.99**	**79.12**	**86.76**	**64.35**	**16.59**	**6.62**	**5.47**	**36,636**	**36,217**	**35,664**	**3,603**	**3,334**	**3,095**	**40,239**	**39,552**	**38,759**

Sequencing centre	Starting DNA	Library protocol	PCR cycles	Sequencing Machine	Illumina chemistry	Reads
UoD	25ng	SureSelect Human all exon V5	8	Illumina Nextseq500	V1	2 × 150bp paired end reads
UoL	750ng	SureSelect Human all exon V5	10	Illumina Hiseq2500	V4	2 × 125bp paired end reads
UoU	750ng	SureSelect Human all exon V5	9	Illumina Hiseq2500	V4	2 × 125bp paired end reads

Stampy/Platypus (pass)	Dundee/Uppsala			Uppsala/Liverpool			Liverpool/Dundee			
File Name	Both	Dundee	Uppsala	Both	Uppsala	Liverpool	Both	Liverpool	Dundee	Between the 3 labs
1-ID9422008	37851	2899	1054	37899	1027	1703	38599	981	2150	37223
2-ID1412384	37457	2735	1342	37774	1030	1903	38396	1267	1787	36879
3-ID1919457	38143	2667	871	37997	1042	1550	38706	823	2111	37470
4-ID1902080	38158	1933	1138	38262	1855	1045	38606	1504	1489	37517
5-ID1919168	37816	1994	1200	38063	969	1506	38342	1205	1462	37301
6-ID1551381	36883	2570	1813	37627	1084	1843	37741	1712	1710	36327
7-ID1413897	37126	3350	966	37139	966	2188	38388	928	2090	36536
8-ID1551357	36965	2995	916	36913	992	1949	37959	882	2004	36360
Average	**37549.875**	**2642.875**	**1162.5**	**37709.25**	**1120.625**	**1710.875**	**38342.125**	**1162.75**	**1850.375**	**36951.63**

91% shared variants between Dundee home pipeline (DP2) and liverpool 1st pipeline

Liverpool (Dundee/Liverpool pipeline)	SNPs			INDELs			Total		
File Name	Both	Dundee	Liverpool	Both	Dundee	Liverpool	Both	Dundee	Liverpool
1-ID9422008	33948	2396	2871	1957	1332	611	35905	3728	3482
2-ID1412384	33893	2502	2896	1976	1327	618	35869	3829	3514
3-ID1919457	33865	2413	2844	1955	1342	558	35820	3755	3402
4-ID1902080	33822	2607	2713	2126	1602	380	35948	4209	3093
5-ID1919168	33808	2470	2850	1954	1367	559	35762	3837	3409
6-ID1551381	33959	2379	2912	1885	1274	647	35844	3653	3559
7-ID1413897	33717	2383	2784	1945	1314	580	35662	3697	3364
8-ID1551357	33070	2507	2809	1938	1376	534	35008	3883	3343
Average	**33760**	**2457**	**2834**	**1967**	**1366**	**560**	**35727**	**3823**	**3395**

➔ ***93% shared SNPs between the 2 different pipelines***
➔ ***60% to 77% shared INDELs between the 2 different pipelines***

98.89% concordancy between genotyping and sequencing files

Stampy/Platypus (Basespace)	Variant number		Genotypes	Shared CHROMxPOS		Same genotypes	
File Name	vcf	gvcf	Affy	With vcf	With gvcf	With vcf	With gvcf
1-ID9422008	43811	48711228	885670	5710	14640	5703	14479
2-ID1412384	44064	48390600	884747	5746	14532	5736	14376
3-ID1919457	43641	48847558	885036	5799	14665	5791	14513
4-ID1902080	43961	49126511	883177	5755	14641	5743	14488
6-ID1551381	43389	47093842	883411	5613	14400	5604	14202
7-ID1413897	44323	48738295	885787	5676	14623	5672	14484
8-ID1551357	42640	48969679	882126	5660	14618	5640	14437
Average	**43320**	**48553959**	**884279**	**5708**	**14588**	**5698**	**14425**

➔2.5x more SNP in common with the genotyping file with gvcf
➔0.2% of genotyping errors (heterozygotes with sequencing / homozygotes with genotyping)

Sequencing in Dundee

Statin

99 cases

39 Godarts

- 28 matching controls
- 5 non matching controls
- 6 GD1 (no DNA)

60 Goshare

- 41 matching controls
- 13 non matching controls
- 6 no ID (waiting for barcode=prochi)

101 controls

All in Godarts

- 81 matching controls (69 + 6GD1 cases + 6 no Goshare ID cases)

+ 8 Super-controls (harmonization of lab)

+ 12 Super-controls

Sequencing in Dundee

ACE/ARB

- **50** cases from Utrecht + controls
- **14** cases from Dundee

➢Sequencing is on going until end of October

Transfer files to Adorial

Ethnic pattern definition

- From Gvcf annotated files (50Mb)
- Comparison to Hapmap database (omni5 chip)

➔List of common variants between hapmap and Gvcf files to extract

➔Addition of our sequencing variants


Universiteit Utrecht

UNIVERSITY OF
LIVERPOOL

UPPSALA
UNIVERSITET


PGXIS PharmacoGenomic Innovative Solutions

Thank you very much for your attention

Thank you to all people on the project
Colin, Moneeza, Myra, Fiona, Andy…