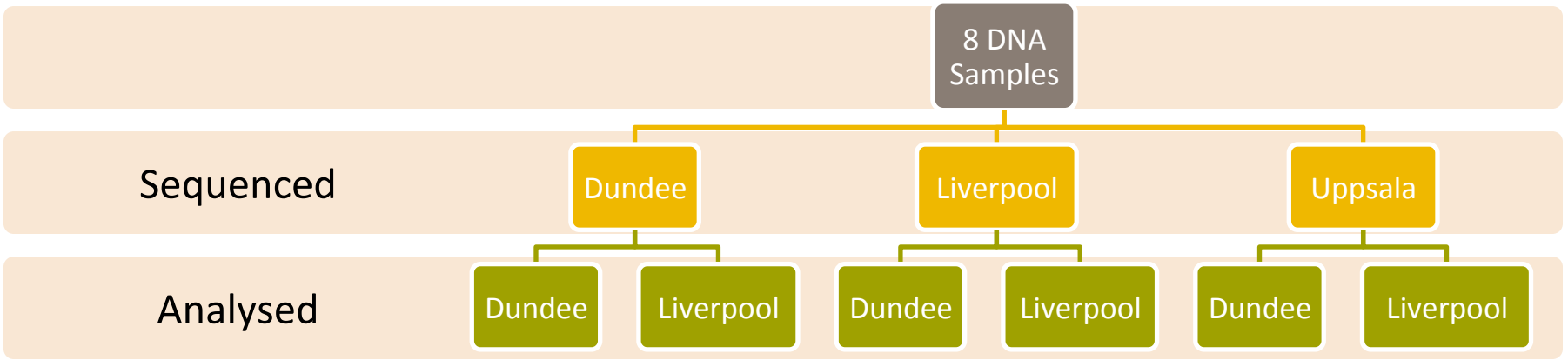
PREDICTION-ADR
WP 3 Sequencing

Kate Bloch
16/09/2015

UNIVERSITY OF LIVERPOOL
THE WOLFSON CENTRE FOR PERSONALISED MEDICINE
MRC Centre for Drug Safety Science

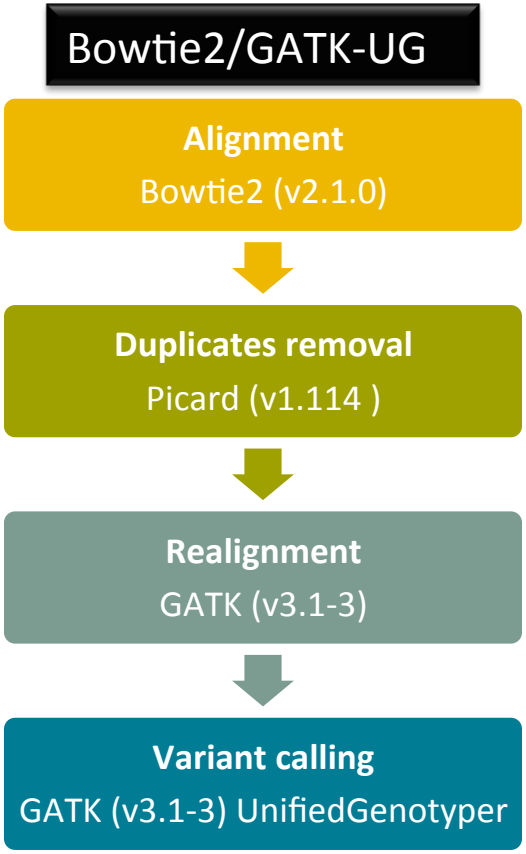
Study design

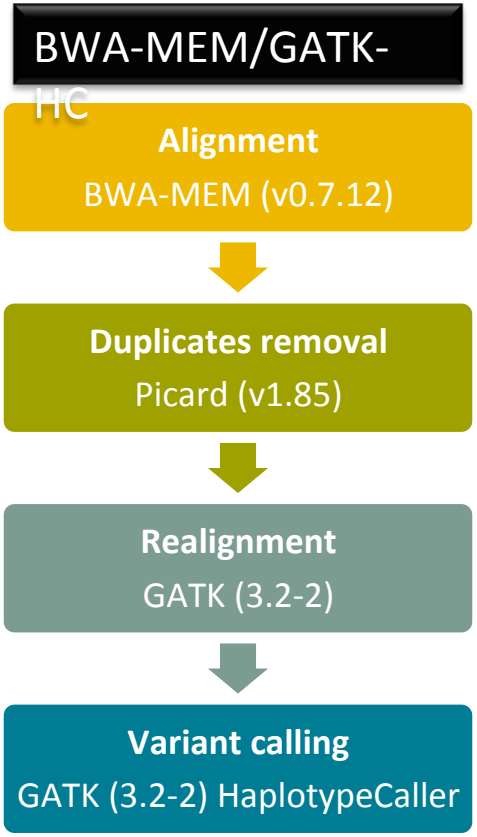
- 8 DNA samples from Dundee University were sent to 3 sequencing centres



Liverpool pipelines

- Two bioinformatics pipelines used:





Liverpool pipelines

- Two bioinformatics pipelines used (Liverpool data):
  - Bowtie2/GATK-UnifiedGenotyper

File Name	% mapped	% at 20x coverage	% on target reads	% duplicates	# SNPs	# INDELs	Total # variants
1-ID9422008	97.68109697	91.59	83.33875027	6.493313114	36,819	2,568	39,387
2-ID1412384	97.71487282	91.35	82.61030604	7.093370604	36,789	2,594	39,383
3-ID1919457	97.78594858	91.54	83.1288109	6.865124988	36,709	2,513	39,222
4-ID1902080	97.39524203	90.72	77.15483542	5.89234598	36,535	2,506	39,041
5-ID1919168	97.72805666	91.71	83.11327615	6.383970281	36,658	2,513	39,171
6-ID1551381	97.77670876	91.53	82.40439841	7.279983982	36,871	2,532	39,403
7-ID1413897	97.77489219	91.36	82.53814636	7.044599796	36,501	2,525	39,026
8-ID1551357	97.6356284	90.77	81.33692059	6.529934088	35,879	2,472	38,351
AVERAGE	**97.6865558**	**91.32**	**81.95**	**6.7**	**36,595**	**2,528**	**39,123**

  - BWA-MEM/GATK-HaplotypeCaller

File Name	% mapped	% at 20x coverage	% on target reads	% duplicates	# SNPs	# INDELs	Total # variants
1-ID9422008	98.2297349	92.35953546	89.71545441	6.49389875	40,328	3,338	43,666
2-ID1412384	98.24218342	92.15313784	89.52921805	7.094950201	41,056	3,463	44,519
3-ID1919457	98.29603169	92.29054641	89.82917822	6.865579844	40,173	3,286	43,459
4-ID1902080	97.90705576	91.48836506	82.53411448	5.893517619	40,510	3,358	43,868
5-ID1919168	98.25276953	92.49803947	89.38008181	6.383893707	40,237	3,333	43,570
6-ID1551381	98.27673116	92.30900024	89.45203518	7.280199027	40,903	3,330	44,233
7-ID1413897	98.27935086	92.12618044	89.34460053	7.045559624	39,956	3,308	43,264
8-ID1551357	98.18309972	91.55810624	87.62314587	6.531063236	40,000	3,281	43,281
AVERAGE	**98.2080224**	**92.1**	**88.43**	**6.7**	**40,395**	**3,337**	**43,733**

GWAS study

- For 7 samples GWAS data (Affy SNP Array 6.0) was received from Dundee University
- The table shows the number (#) of variants present in each file

Sample	PRO CHI ID	# 0/0	# 0/1	# 1/1	# ./.	# chr X	# chr Y	# chr MT	Total	Total – (./.)	Total – MT, X, Y & (./.)
1	akh1040564	479,003	230,431	176,527	604	34,875	608	289	**886,565**	885,961	**850,189**
2	akh0204487	479,326	231,676	174,036	1,527	34,796	611	286	**886,565**	885,038	**849,345**
3*	akh2521638	649,927*	235,400	0*	1,238	34,894	45	289	**886,565**	885,327	**850,099**
4	akh5712067	472,570	238,264	172,633	3,098	34,829	45	288	**886,565**	883,467	**848,305**
6	akh1649529	475,389	234,409	173,905	2,862	34,799	45	290	**886,565**	883,703	**848,569**
7	akh5554834	478,961	231,782	175,335	487	34,920	612	289	**886,565**	886,078	**850,257**
8	akh1348973	475,709	230,546	176,160	4,150	34,501	607	287	**886,565**	882,415	**847,020**
Average		**501,555**	**233,215**	**149,799**	**1,995**	**34,802**	**368**	**288**	**886,565**	**884,570**	**849,112**

* QC ?

GWAS-bed-exome comparison

- The table shows # of overlapping variants between GWAS –bed – exome (Liverpool data)
- Comparison of 2 Liverpool pipelines

GWAS file with 0/0 with MT, X, Y			bed file (with MT, X, Y)	Overlap with GWAS	
Sample	**PRO CHI ID**	**# variants**	**Overlap with bed**	**Bowtie2/GATK-UG**	**BWA-MEM/GATK-HC**
1	akh1040564	885,961	14,149	5,522	5,698
2	akh0204487	885,038	14,137	5,553	5,749
3	akh2521638	885,327	14,153	5,539	5,726
4	akh5712067	883,467	14,121	5,469	5,655
6	akh1649529	883,703	14,119	5,486	5,683
7	akh5554834	886,078	14,151	5,488	5,652
8	akh1348973	882,415	14,092	5,414	5,599
Average		**884,570**	**14,132 (1.6% of GWAS)**	**5,496 (38.9% with bed)**	**5,680 (40.19% with bed)**

- On average **184 more** variants overlapping with GWAS were detected using **BWA/GATK-HC**
- No overlapping variants on chr X and Y

GWAS-bed-exome comparison

➢The table shows the number of overlapping variants between GWAS (w/o 0/0, with X, Y, MT) –bed – Liverpool data

GWAS file without 0/0 with MT,X,Y			bed file (with X, Y, MT)	Overlap with GWAS	
Sample	PRO CHI ID	# variants	GWAS overlap with bed	Bowtie2/GATK-UG	BWA-MEM/GATK-HC
1	akh1040564	406,958	5,994	4,134	4,270
2	akh0204487	405,712	5,996	4,150	4,300
3	akh2521638	235,400*	3,458*	3,292*	3,390*
4	akh5712067	410,897	5,969	4,056	4,203
6	akh1649529	408,314	5,920	4,065	4,209
7	akh5554834	407,117	5,926	4,074	4,205
8	akh1348973	406,706	5,897	3,999	4,133
Average		**383,015**	**5,594 (1.47% of GWAS)**	**3,967 (72.4% with bed)**	**4,101 (73.3% with bed)**

➢ On average **134 more** overlapping variants with GWAS were detected using **BWA/GATK-HC**

➢No overlapping variants on chr X and Y

Dundee and Uppsala data

- BWA-MEM/GATK-HC pipeline showed superior performance with Uppsala and Dundee data as well

- Dundee data

Average	% mapped	% at 20x coverage	% on target reads	% duplicates	# SNPs	# INDELs	Total # variants
Bowtie2/GATK-UG	95.73904092	93.36	66.02416661	16.53811027	38,173	2,795	40,968
BWA-MEM/GATK-HC	**95.78102805**	**94.14881755**	**66.3344713**	**16.59655691**	**41,467**	**3,567**	**45,034**

- Uppsala data

Average	% mapped	% at 20x coverage	% on target reads	% duplicates	# SNPs	# INDELs	Total # variants
Bowtie2/GATK-UG	95.39741884	87.11875	59.48010448	5.431493278	36,493	2,688	39,181
BWA-MEM/GATK-HC	**96.66199013**	**87.87995131**	**60.02636039**	**5.462527691**	**40,243**	**3,413**	**43,656**

Intersection Dundee-Liverpool-Uppsala

- Table shows overlap of variants between sequencing centres
- Dundee more variants detected (more reads)
- Liverpool more (on average) overlapping variants with GWAS

Sample	Unique			Shared				Total			Overlap with GWAS		
	Dun	Liv	Upp	Dundee-Liverpool-Uppsala	Dundee-Liverpool	Dundee-Uppsala	Liverpool-Uppsala	Dun	Liv	Upp	Dun	Liv	Upp
1-ID9422008	**2,040**	274	420	41,883	743	420	766	**45,086**	43,666	43,489	5,683	**5,698**	5,688
2-ID1412384	**1,851**	337	502	42,268	958	443	956	**45,520**	44,519	44,169	5,722	**5,749**	5,726
3-ID1919457	**2,155**	263	424	42,032	640	602	524	**45,429**	43,459	43,582	**5,731**	5,726	5,730
4-ID1902080	**1,831**	272	557	42,472	532	638	592	**45,473**	43,868	44,259	**5,660**	5,655	5,675
5-ID1919168	**1,751**	270	472	42,073	559	545	668	**44,928**	43,570	43,758	n/a	n/a	n/a
6-ID1551381	**1,377**	395	572	41,731	772	380	1,335	**44,260**	44,233	44,018	5,614	**5,683**	5,665
7-ID1413897	**2,094**	256	448	41,308	997	426	703	**44,825**	43,264	42,885	5,635	**5,652**	5,624
8-ID1551357	**1,930**	281	433	41,485	841	494	674	**44,750**	43,281	43,086	**5,599**	**5,599**	5,607
Average	**1,879**	294	479	41,907	755	494	777	**45,034**	43,733	43,656	5,663	**5,680**	5,674

UNIVERSITY OF LIVERPOOL

THE WOLFSON CENTRE FOR PERSONALISED MEDICINE

MRC Centre for Drug Safety Science

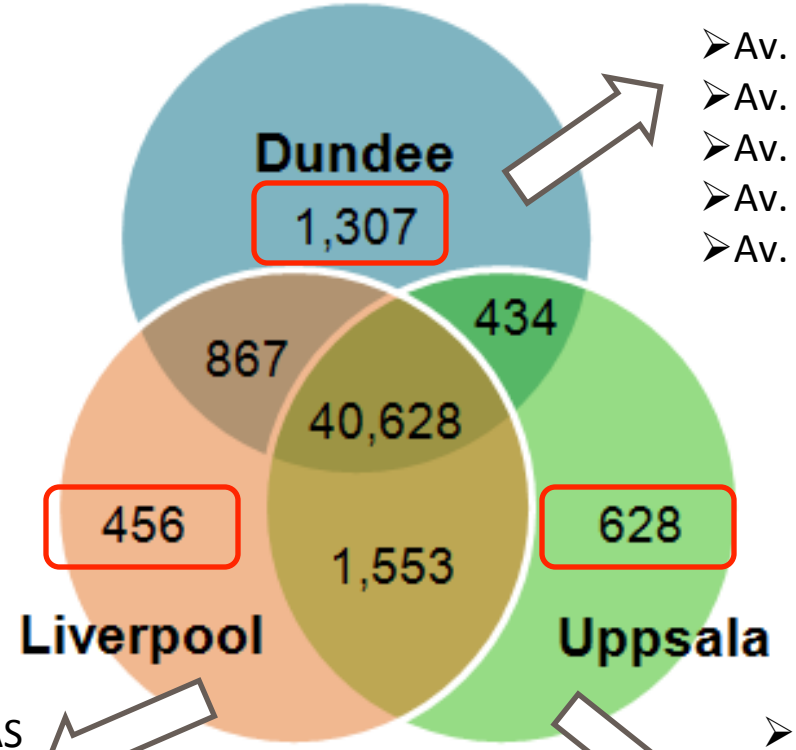
Comparing sequencing centres

- Sub-sampling 40M reads (randomly sub-sampled from quality and adapter trimmed)
- BWA/MEM-GATK-HC

Sequencing centre	Total # of reads	Total sub-sampled # of reads	% Reads uniquely mapped (w/ duplicates)	% Reads uniquely mapped (w/o duplicates)	% Number of bases in the capture regions with coverage more than 20x	Total # of variants	Overlap with GWAS (%)
Dundee	**93,956,352**	40,000,000	81.27600781	73.16451688	87.26100007	43,236	5,587 (12.9%)
Liverpool	50,097,759	40,000,000	**88.44795688**	**83.5794075**	**89.14402782**	**43,504**	**5,667** (13%)
Uppsala	50,022,238	40,000,000	63.54013094	60.61416813	86.33650113	43,243	5,649 **(13.1%)**

Comparing sequencing centres

- Sub-sampling 40M reads (randomly sub-sampled from quality and adapter trimmed)
- Overlap of variants detected in Dundee, Liverpool and Uppsala (average)



- Av. 0.29 % overlap with GWAS
- Av. 28% in dbSNP
- Av. Ts/Tv 0.25
- Av. Hom 8.5%
- Av. Het 91.5%

- Av. 1.32% overlap with GWAS
- Av. 62.1% in dbSNP
- Av. Ts/Tv 1.45
- Av. Hom 16%
- Av. Het 84%

- Av. **2.9%** overlap with GWAS
- Av. **67.1%** in dbSNP
- Av. Ts/Tv 1.24
- Av. Hom 20%
- Av. Het 80%