

Introduction to **GWAS**

Common Data Types and Formats

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)



HerrFalloppio

Oscar González-Recio

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)



OscarGenomics



Genotyping and Sequencing

A very brief overview

The first steps – Biomarkers

A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. I. THE NUMBER OF ALLELES AT DIFFERENT LOCI IN *DROSOPHILA PSEUDOBSCURA*¹

J. L. HUBBY AND R. C. LEWONTIN

Department of Zoology, University of Chicago, Chicago, Illinois

Received March 30, 1966

¹ The work reported here was supported in part by grants from the National Science Foundation (GB 3013) and the Public Health Service (5M-13206).

Genetics 84: 517-524 August 1968.



D. pseudoobscura (F)

582

J. L. HUBBY AND R. C. LEWONTIN

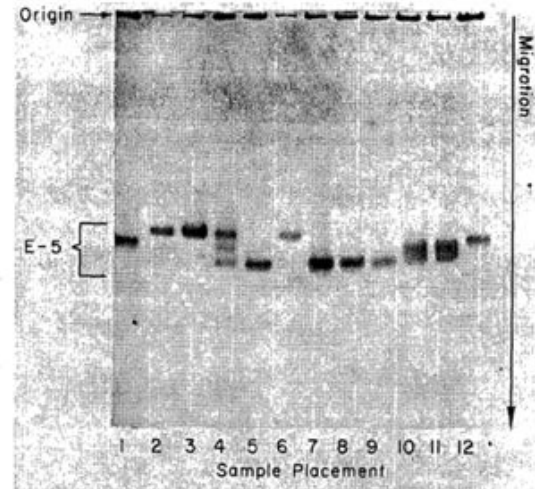


FIGURE 1.—Gel illustrating sample placement and typical results of strain analysis for Esterase-5. The first and the last samples were derived from the standard reference strain (E-5^{1.00}), while positions 2 through 6 were obtained from five individuals of one strain and positions 7 through 11 are from five individuals of a second strain. Positions 2, 3, and 6 contain Esterase-5⁹⁵, position 5 contains Esterase-5^{1.12}, and position 4 contains Esterase-5⁹⁵, Esterase-5^{1.12}, and a site of activity between them. Positions 7, 8, and 9 contain Esterase-5^{1.12} and positions 10 and 11 contain Esterase-5^{1.00} and Esterase-5^{1.12}. A site of activity midway between the latter two is barely discernible. In all the figures the direction of migration of the protein is down toward the anode.

From few to many markers – Molecular markers (DNA markers)

- arise from **different classes of DNA mutations** such as substitution mutations (point mutations), rearrangements (insertions or deletions) or errors in replication of tandemly repeated DNA
- are usually located in **non-coding regions** of DNA
- are practically **unlimited in number** and are **not affected by environmental factors** and/or the developmental stage of the plant
- **RFLP, AFLP, RAPD, SSR (microsatellites), SNP**

From few to many markers – Molecular markers (DNA markers)



Euphytica (2005) 142: 169–196
DOI: 10.1007/s10681-005-1681-5

© Springer 2005

An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts

B.C.Y. Collard^{1,4,*}, M.Z.Z. Jahufer², J.B. Brouwer³ & E.C.K. Pang¹

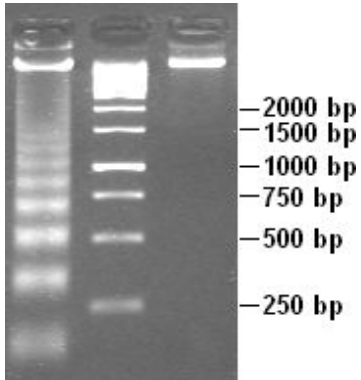
¹*Department of Biotechnology and Environmental Biology, RMIT University, P.O. Box 71, Bundoora, Victoria 3083, Australia;* ²*AgResearch Ltd., Grasslands Research Centre, Tennent Drive, Private Bag 11008, Palmerston North, New Zealand;* ³*P.O. Box 910, Horsham, Victoria, Australia 3402;* ⁴*Present address: Plant Breeding, Genetics and Biotechnology Division, International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines;* (*author for correspondence: e-mail: bcycollard@hotmail.com)

Received 11 July 2004; accepted 2 February 2005

Genotyping Systems

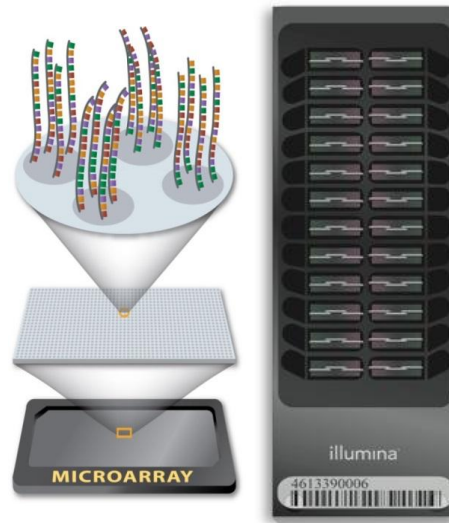
Marker gel

(a few markers)



SNP array (or GBS)

(100s -1,000,000s)



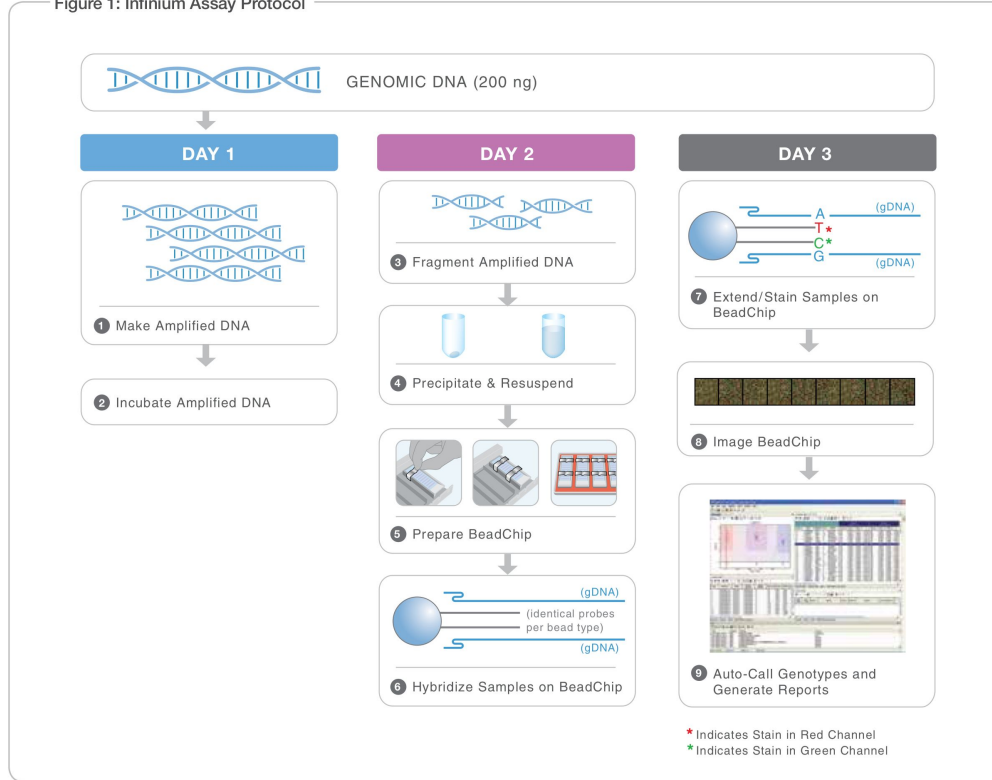
Genome sequencer

(1,000,000s +)

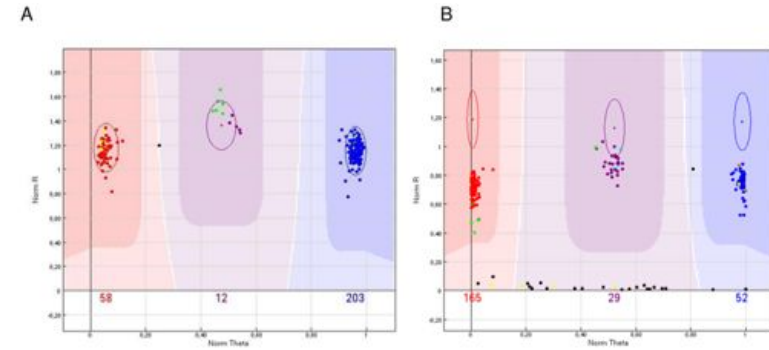


SNP array genotyping

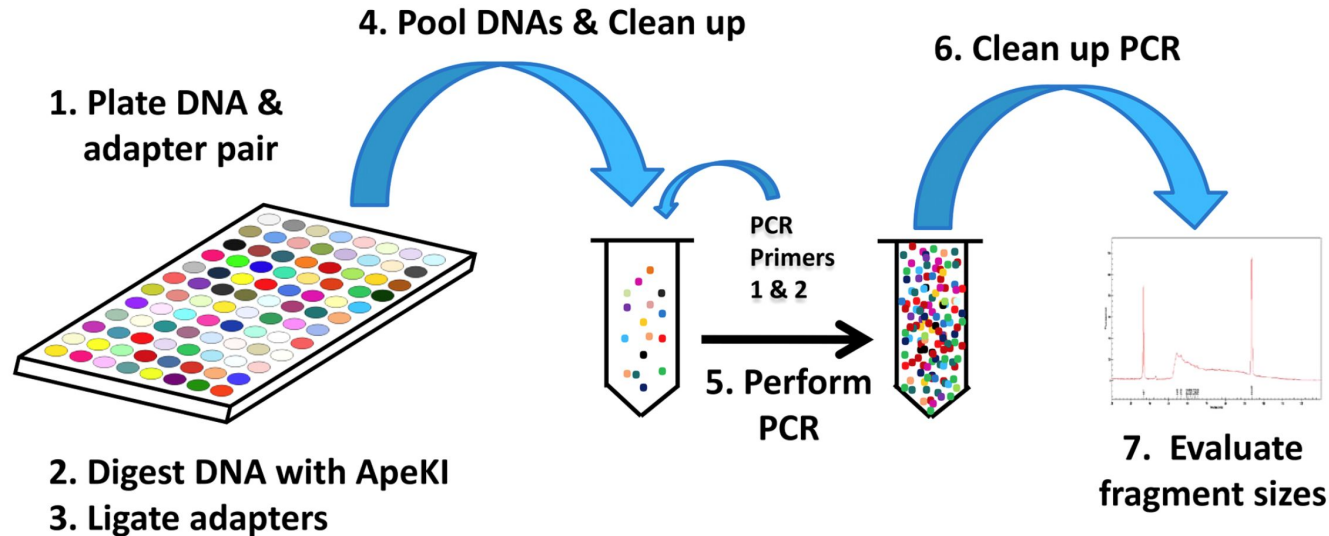
Figure 1: Infinium Assay Protocol



Genotype calling
3 genotypes: AA, AG, GG



Reduced representation sequencing – Genotyping-by-Sequencing (GBS)

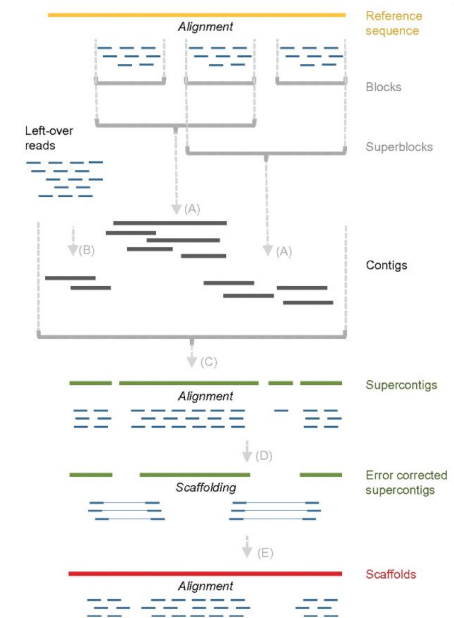


doi: <https://doi.org/10.1371/journal.pone.0019379.g002>

The Next Generation Sequencing Revolution



Hmmm...now the data is here....so what now?

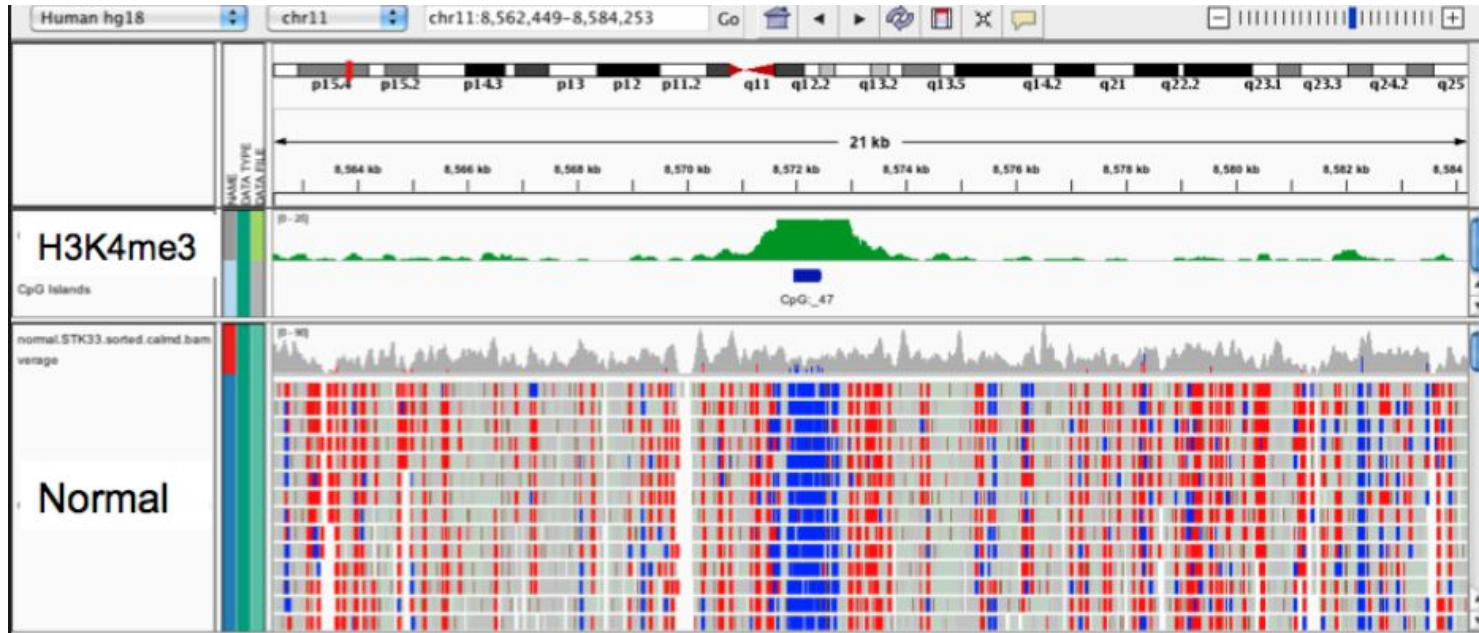


Assembly!



The Next Generation Sequencing Revolution

Millions of polymorphisms in the genome sequences...



Genotyping formats

Two major genotyping formats (but there are many...)

Plink format ped/map files

<https://www.cog-genomics.org/plink2/>

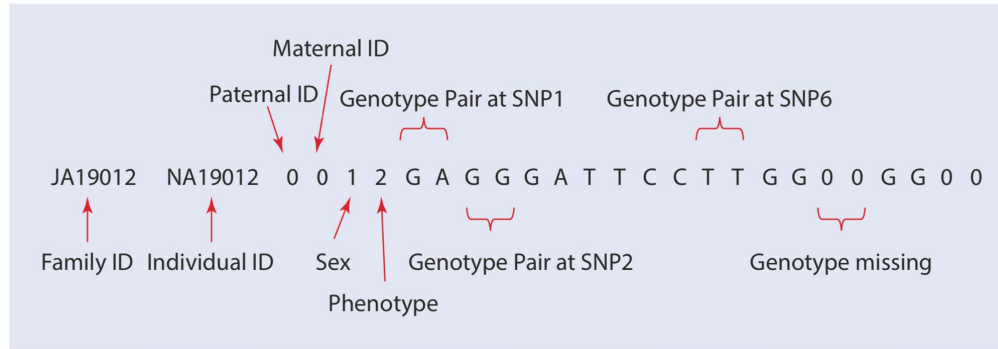
Variant calling format vcf files

<https://www.internationalgenome.org/wiki/Analysis/vcf4.0/>

<https://vcftools.github.io/index.html>



PLINK - the .ped file



Column 1 = Family ID
 Column 2 = Individual ID
 Column 3 = Paternal ID (zero for missing)
 Column 4 = Maternal ID (zero for missing)
 Column 5 = Sex
 Column 6 = Phenotype (1=unaffected, 2=affected, and 0=missing)
 Column 7, 8 = genotype pair of the first SNP1 (zero for missing)
 Column 9, 10 = genotype pair of the second SNP2 (zero for means missing)
 ...
 Column 457393, 457394 = genotype pair of the last SNP228694

Kim J.H. (2019) GWAS Data Analysis. In: Genome Data Analysis. Learning Materials in Biosciences. Springer, Singapore

PLINK - the .map file

		genotype.ped																							
JA19012	NA19012	0	0	1	2	G	A	G	G	G	A	T	T	C	C	T	T	G	G	0	0	G	G	0	0
		1	rs6681049	0	789870																				
		1	rs4074137	0	1016570																				
		1	rs7540009	0	1050098																				
		1	rs1891905	0	1090080																				
		1	rs9729550	0	1125105																				
		1	rs3813196	0	1159244																				
		1	rs6704013	0	1187454																				

genotype.map

- Column 1 = chromosome number
- Column 2 = SNP ID
- Column 3 = Genetic Distance (morgans)
- Column 4 = physical base-pair position (bp)

Kim J.H. (2019) GWAS Data Analysis. In: Genome Data Analysis. Learning Materials in Biosciences. Springer, Singapore



Variant calling format - the .vcf file

Data info

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:CQ:DP:HQ 0/0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:CQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:CQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:CQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234567 microsat1 CTC G,CTCT 50 PASS NS=3;DP=9;AA=G GT:CQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

CHROM → chromosome/contig

POS → position on chr/contig

ID → SNP name/ID

REF → reference genome allele

ALT → alternative allele (. / 1 or more)

QUAL → Phred scaled quality

-10log10 (call in Alt is wrong)

E.g. 1/10 chance of mistake → 10

FILTER ☐ quality filter (q10 → quality < 10)

INFO ☐ further information

0/1 is not the same as 0|1 !! (unphased / phased)

