Chapter 9:
Analysis of next-generation sequence data

(this presentation is modified version of the original)

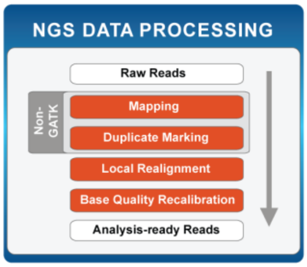
Jonathan Pevsner, Ph.D.
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(Wiley-Liss, 3rd edition, 2015)
You may use this PowerPoint for teaching purposes

Outline:
Analysis of Next-Generation Sequence (NGS) Data

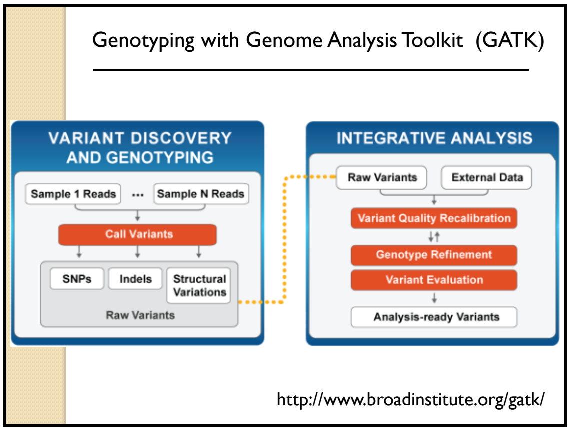
Introduction
DNA sequencing technologies
  Sanger sequencing; NGS; Illumina; pyrosequencing;
  ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics
Analysis of NGS sequencing of genomic DNA
  Overview
  Topic 1: Design
  Topic 2: FASTQ
  Topic 3: Assembly
  Topic 4: Alignment
  Topic 5: SAM/BAM
  Topic 6: Variant calling: SNVs
  Topic 7: Variant calling: SVs
  Topic 8: VCF
  Topic 9: Visualizing NGS data
  Topic 10: Significance
Specialized applications of NGS
Perspective

Genotyping with Genome Analysis Toolkit (GATK)

Popular suite of tools used for genotyping and variant discovery


NGS DATA PROCESSING
Raw Reads
Non-GATK
Mapping
Duplicate Marking
Local Realignment
Base Quality Recalibration
Analysis-ready Reads

http://www.broadinstitute.org/gatk/



Genotyping with Genome Analysis Toolkit (GATK)

VARIANT DISCOVERY AND GENOTYPING

Sample 1 Reads ... Sample N Reads

Call Variants

SNPs

Indels

Structural Variations

Raw Variants

INTEGRATIVE ANALYSIS

Raw Variants

External Data

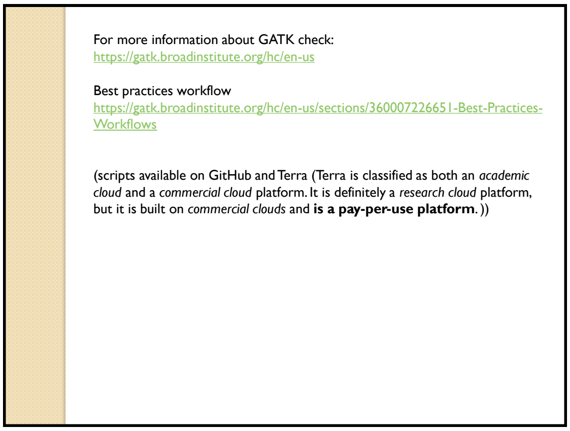
Variant Quality Recalibration

Genotype Refinement

Variant Evaluation

Analysis-ready Variants

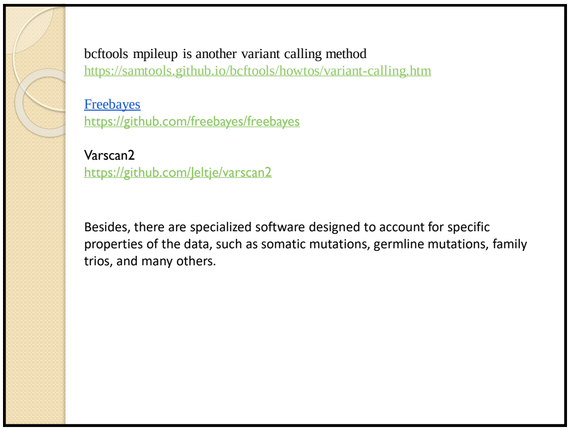
http://www.broadinstitute.org/gatk/



For more information about GATK check:
https://gatk.broadinstitute.org/hc/en-us

Best practices workflow
https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows

(scripts available on GitHub and Terra (Terra is classified as both an *academic cloud* and a *commercial cloud* platform. It is definitely a *research cloud* platform, but it is built on *commercial clouds* and **is a pay-per-use platform**.))

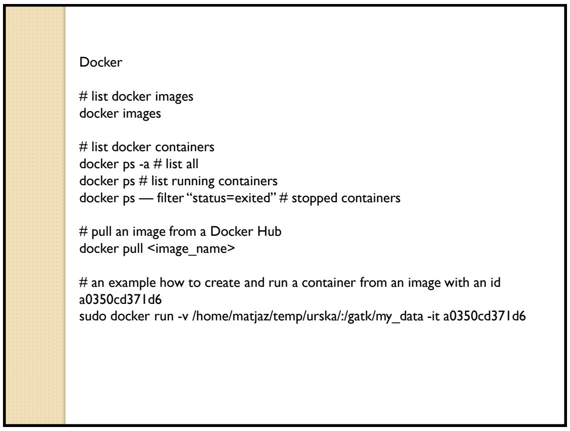


bcftools mpileup is another variant calling method
https://samtools.github.io/bcftools/howtos/variant-calling.htm

Freebayes
https://github.com/freebayes/freebayes

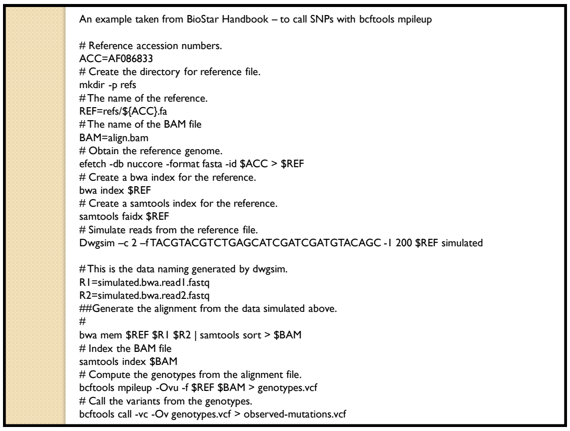
Varscan2
https://github.com/Jeltje/varscan2

Besides, there are specialized software designed to account for specific properties of the data, such as somatic mutations, germline mutations, family trios, and many others.



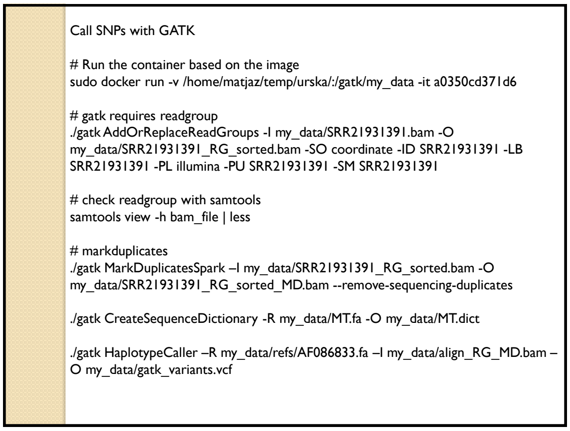
Docker

```
# list docker images
docker images

# list docker containers
docker ps -a # list all
docker ps # list running containers
docker ps — filter “status=exited” # stopped containers

# pull an image from a Docker Hub
docker pull <image_name>

# an example how to create and run a container from an image with an id
a0350cd371d6
sudo docker run -v /home/matjaz/temp/urska/:/gatk/my_data -it a0350cd371d6
```



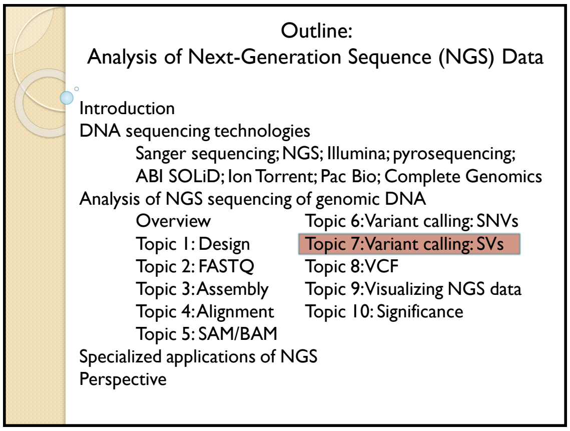
An example taken from BioStar Handbook – to call SNPs with bcftools mpileup

```
# Reference accession numbers.
ACC=AF086833
# Create the directory for reference file.
mkdir -p refs
# The name of the reference.
REF=refs/${ACC}.fa
# The name of the BAM file
BAM=align.bam
# Obtain the reference genome.
efetch -db nuccore -format fasta -id $ACC > $REF
# Create a bwa index for the reference.
bwa index $REF
# Create a samtools index for the reference.
samtools faidx $REF
# Simulate reads from the reference file.
Dwgsim –c 2 –f TACGTACGTCTGAGCATCGATCGATGTACAGC -1 200 $REF simulated

# This is the data naming generated by dwgsim.
R1=simulated.bwa.read1.fastq
R2=simulated.bwa.read2.fastq
##Generate the alignment from the data simulated above.
#
bwa mem $REF $R1 $R2 | samtools sort > $BAM
# Index the BAM file
samtools index $BAM
# Compute the genotypes from the alignment file.
bcftools mpileup -Ovu -f $REF $BAM > genotypes.vcf
# Call the variants from the genotypes.
bcftools call -vc -Ov genotypes.vcf > observed-mutations.vcf
```



Call SNPs with GATK

```
# Run the container based on the image
sudo docker run -v /home/matjaz/temp/urska/:/gatk/my_data -it a0350cd371d6

# gatk requires readgroup
./gatk AddOrReplaceReadGroups -I my_data/SRR21931391.bam -O
my_data/SRR21931391_RG_sorted.bam -SO coordinate -ID SRR21931391 -LB
SRR21931391 -PL illumina -PU SRR21931391 -SM SRR21931391

# check readgroup with samtools
samtools view -h bam_file | less

# markduplicates
./gatk MarkDuplicatesSpark –I my_data/SRR21931391_RG_sorted.bam -O
my_data/SRR21931391_RG_sorted_MD.bam --remove-sequencing-duplicates

./gatk CreateSequenceDictionary -R my_data/MT.fa -O my_data/MT.dict

./gatk HaplotypeCaller –R my_data/refs/AF086833.fa –I my_data/align_RG_MD.bam –
O my_data/gatk_variants.vcf
```

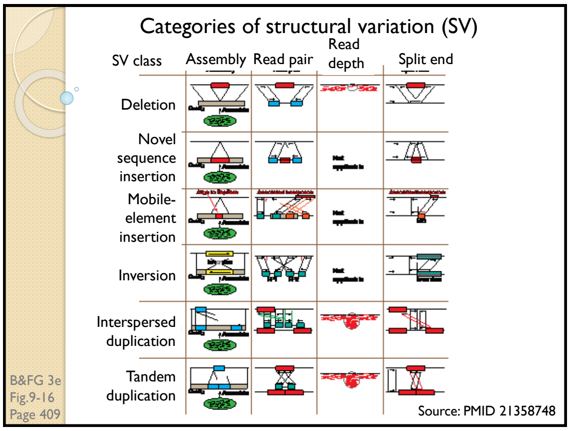


Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction
DNA sequencing technologies
Sanger sequencing; NGS; Illumina; pyrosequencing;
ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics
Analysis of NGS sequencing of genomic DNA

Overview	Topic 6: Variant calling: SNVs
Topic 1: Design	Topic 7: Variant calling: SVs
Topic 2: FASTQ	Topic 8: VCF
Topic 3: Assembly	Topic 9: Visualizing NGS data
Topic 4: Alignment	Topic 10: Significance
Topic 5: SAM/BAM	

Specialized applications of NGS
Perspective

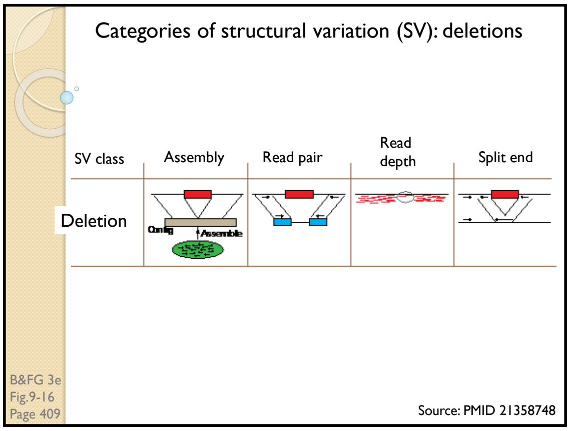


Categories of structural variation (SV)

SV class	Assembly	Read pair	Read depth	Split end
Deletion				
Novel sequence insertion			Not applicable	
Mobile-element insertion			Not applicable	
Inversion			Not applicable	
Interspersed duplication				
Tandem duplication				

B&FG 3e
Fig.9-16
Page 409

Source: PMID 21358748

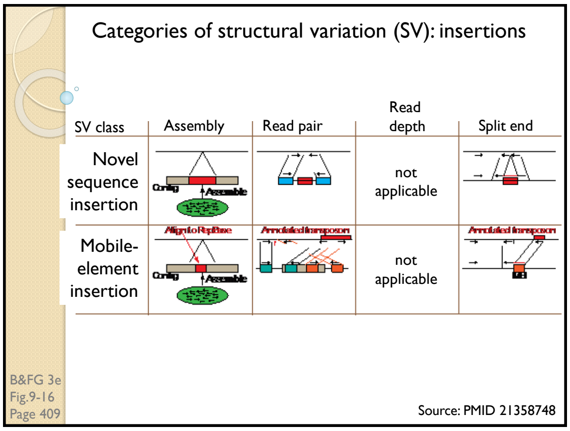


Categories of structural variation (SV): deletions

SV class	Assembly	Read pair	Read depth	Split end
Deletion				

B&FG 3e
Fig.9-16
Page 409

Source: PMID 21358748

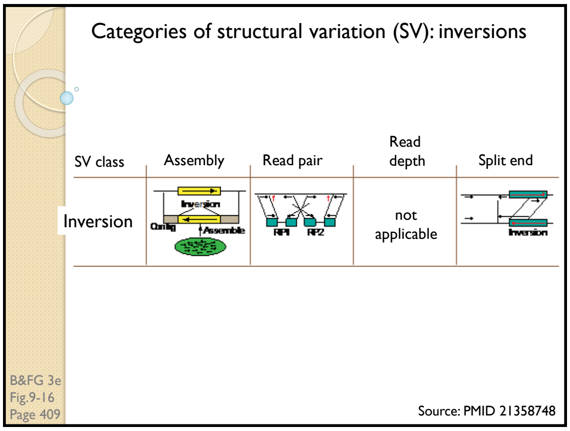


Categories of structural variation (SV): insertions

SV class	Assembly	Read pair	Read depth	Split end
Novel sequence insertion			not applicable	
Mobile-element insertion			not applicable	

B&FG 3e
Fig.9-16
Page 409

Source: PMID 21358748

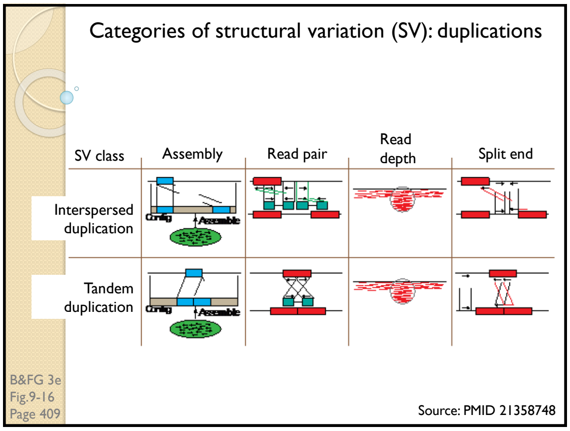


Categories of structural variation (SV): inversions

SV class	Assembly	Read pair	Read depth	Split end
Inversion			not applicable	

B&FG 3e
Fig.9-16
Page 409

Source: PMID 21358748



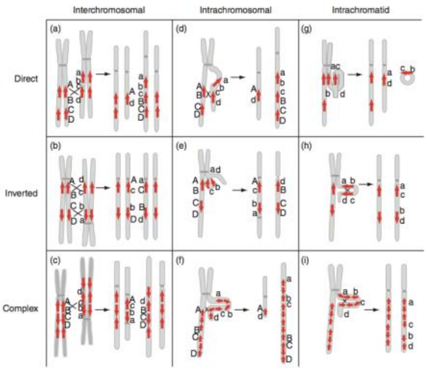
Categories of structural variation (SV): duplications

SV class	Assembly	Read pair	Read depth	Split end
Interspersed duplication				
Tandem duplication				

B&FG 3e
Fig.9-16
Page 409

Source: PMID 21358748

Mechanisms of creating genomic rearrangements



B&FG 3e
Fig. 8.19
Page 352

Explanation of Fig 8.19

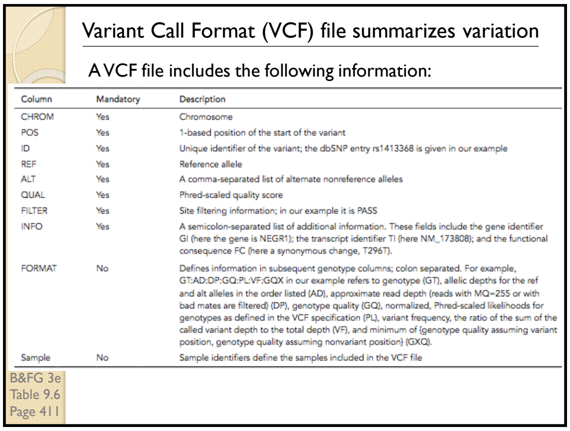
Mechanisms of creating genomic rearrangements. Non-allelic homologous recombination (NAHR) based on low-copy repeats (LCRs) or segmental duplications cause these changes. The orientation of the LCRs may be **head-to-head (top row), head-to-toe (middle row**), or complex (bottom row) involving DNA exchanges that are interchromosomal (left column), intrachromosomal (middle column), or intrachromatid (right column). For each of the nine scenarios the chromosomal configuration is shown as well as the products of unequal crossing over. (a) Unequal cross-overs between directly ordered repeats lead to a duplication and a deletion. (b) Mechanism of forming an inversion. (c) Interchromosomal exchange between inverted repeats causes inversions and can result in duplications and deletions. (d) Mispairing of direct repeats leads to an intrachromosomal deletion/duplication. (e) An inversion results from intrachromosomal unequal exchange between inverted repeats. (f) Complex repeats lead to an intrachromosomal deletion/duplication. (g) A deletion and an acentric fragment result from intrachromatid mispairing due to direct low-copy repeats. (h) An intrachromatid loop of inverted repeats results in an inversion. (i) Complex repeats lead to intrachromatid mispairing and an inversion. Redrawn from Stankiewicz and Lupski (2002) with permission from Elsevier.

Outline:
Analysis of Next-Generation Sequence (NGS) Data

Introduction
DNA sequencing technologies
- Sanger sequencing; NGS; Illumina; pyrosequencing; ABI SOLiD; Ion Torrent; Pac Bio; Complete Genomics

Analysis of NGS sequencing of genomic DNA
- Overview
- Topic 1: Design
- Topic 2: FASTQ
- Topic 3: Assembly
- Topic 4: Alignment
- Topic 5: SAM/BAM
- Topic 6: Variant calling: SNVs
- Topic 7: Variant calling: SVs
- Topic 8: VCF
- Topic 9: Visualizing NGS data
- Topic 10: Significance

Specialized applications of NGS
Perspective

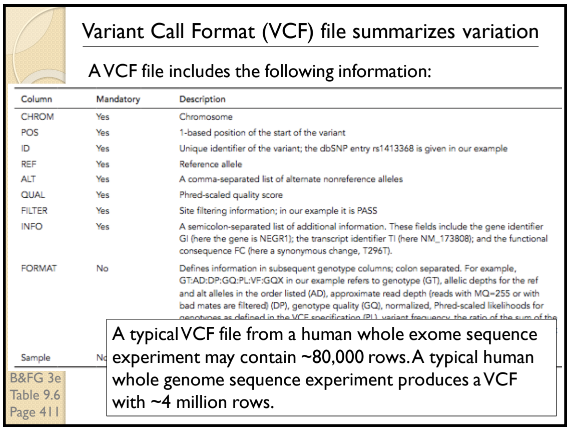


Variant Call Format (VCF) file summarizes variation

A VCF file includes the following information:

Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GQX in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification (PL), variant frequency, the ratio of the sum of the called variant depth to the total depth (VF), and minimum of {genotype quality assuming variant position, genotype quality assuming nonvariant position} (GXQ).
Sample	No	Sample identifiers define the samples included in the VCF file

B&FG 3e
Table 9.6
Page 411



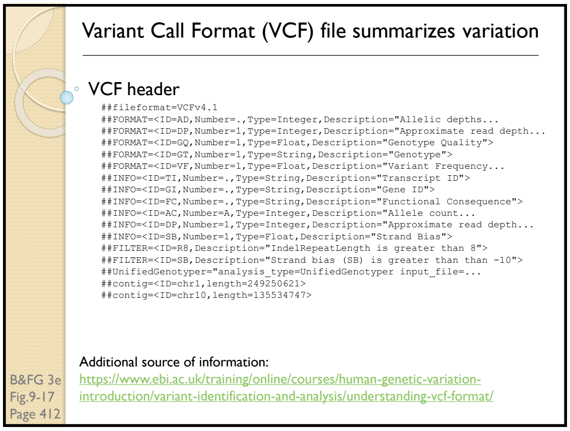
Variant Call Format (VCF) file summarizes variation

A VCF file includes the following information:

Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GQX in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for
Sample	N	

B&FG 3e
Table 9.6
Page 411

A typical VCF file from a human whole exome sequence experiment may contain ~80,000 rows. A typical human whole genome sequence experiment produces a VCF with ~4 million rows.



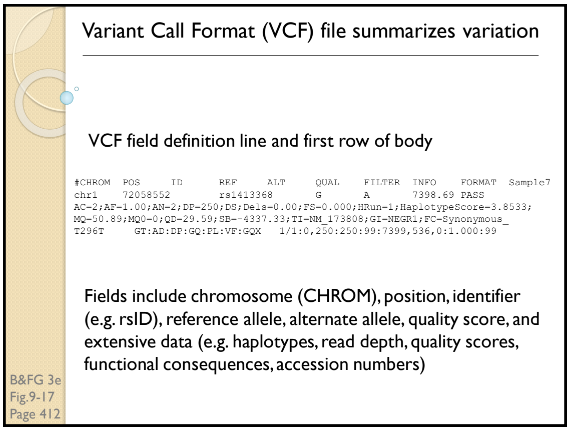
Variant Call Format (VCF) file summarizes variation

- VCF header

```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths...
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth...
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=VF,Number=1,Type=Float,Description="Variant Frequency...
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
##INFO=<ID=GI,Number=.,Type=String,Description="Gene ID">
##INFO=<ID=FC,Number=.,Type=String,Description="Functional Consequence">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth...
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##FILTER=<ID=R8,Description="IndelRepeatLength is greater than 8">
##FILTER=<ID=SB,Description="Strand bias (SB) is greater than than -10">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=...
##contig=<ID=chr1,length=249250621>
##contig=<ID=chr10,length=135534747>
```

Additional source of information:
https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/understanding-vcf-format/

B&FG 3e
Fig.9-17
Page 412



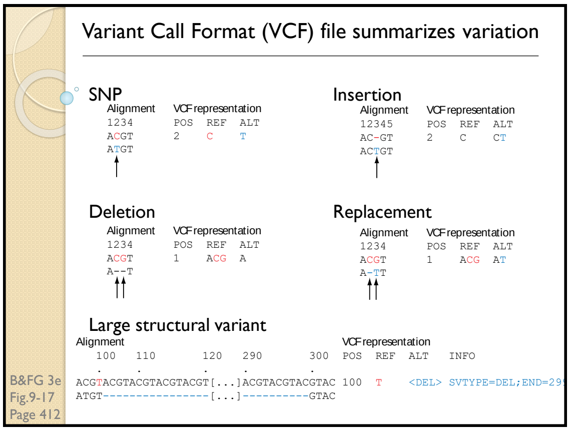
Variant Call Format (VCF) file summarizes variation

VCF field definition line and first row of body

```
#CHROM  POS      ID       REF      ALT      QUAL     FILTER  INFO     FORMAT  Sample7
chr1    72058552          rs1413368         G        A       7398.69 PASS
AC=2;AF=1.00;AN=2;DP=250;DS;Dels=0.00;FS=0.000;HRun=1;HaplotypeScore=3.8533;
MQ=50.89;MQ0=0;QD=29.59;SB=-4337.33;TI=NM_173808;GI=NEGR1;FC=Synonymous_
T296T     GT:AD:DP:GQ:PL:VF:GQX   1/1:0,250:250:99:7399,536,0:1.000:99
```

Fields include chromosome (CHROM), position, identifier (e.g. rsID), reference allele, alternate allele, quality score, and extensive data (e.g. haplotypes, read depth, quality scores, functional consequences, accession numbers)

B&FG 3e
Fig.9-17
Page 412



Variant Call Format (VCF) file summarizes variation

SNP

```
Alignment   VCF representation
1234        POS  REF  ALT
ACGT        2    C    T
ATGT
```

Insertion

```
Alignment   VCF representation
12345       POS  REF  ALT
AC-GT       2    C    CT
ACTGT
```

Deletion

```
Alignment   VCF representation
1234        POS  REF  ALT
ACGT        1    ACG  A
A--T
```

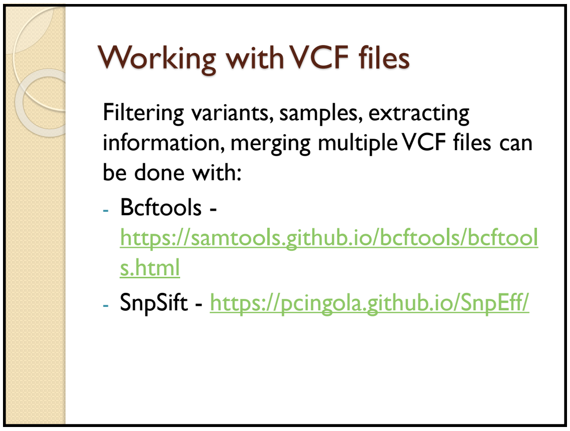
Replacement

```
Alignment   VCF representation
1234        POS  REF  ALT
ACGT        1    ACG  AT
A-TT
```

Large structural variant

```
Alignment                                        VCF representation
   100    110          120    290          300   POS  REF  ALT   INFO
    .      .            .      .            .
ACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC 100  T    <DEL> SVTYPE=DEL;END=29
ATGT----------------[...]----------GTAC
```

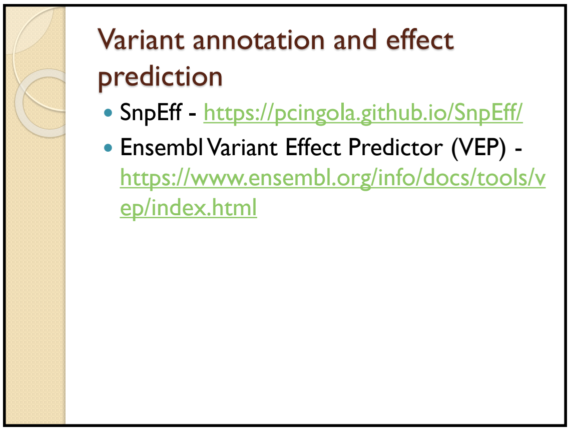
B&FG 3e
Fig.9-17
Page 412



Working with VCF files

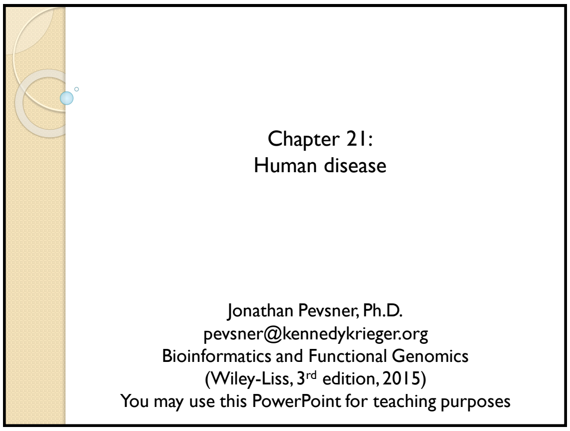
Filtering variants, samples, extracting information, merging multiple VCF files can be done with:

- Bcftools - https://samtools.github.io/bcftools/bcftools.html
- SnpSift - https://pcingola.github.io/SnpEff/



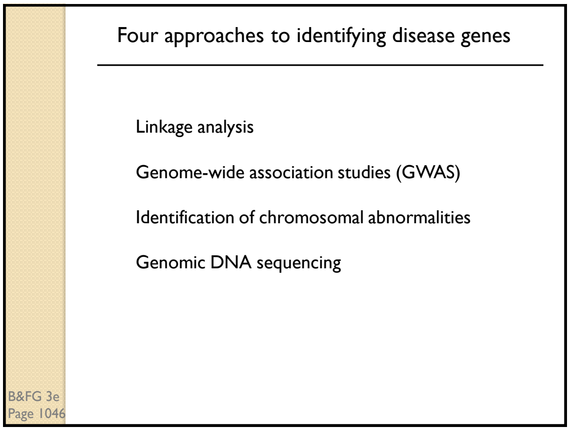
Variant annotation and effect prediction

- SnpEff - https://pcingola.github.io/SnpEff/
- Ensembl Variant Effect Predictor (VEP) - https://www.ensembl.org/info/docs/tools/vep/index.html



Chapter 21:
Human disease

Jonathan Pevsner, Ph.D.
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(Wiley-Liss, 3rd edition, 2015)
You may use this PowerPoint for teaching purposes



Four approaches to identifying disease genes

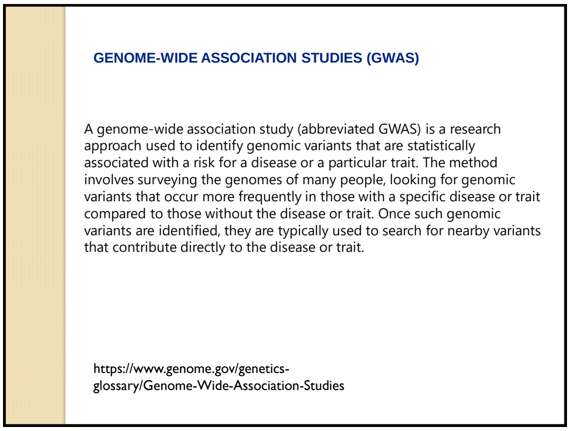
Linkage analysis

Genome-wide association studies (GWAS)

Identification of chromosomal abnormalities

Genomic DNA sequencing

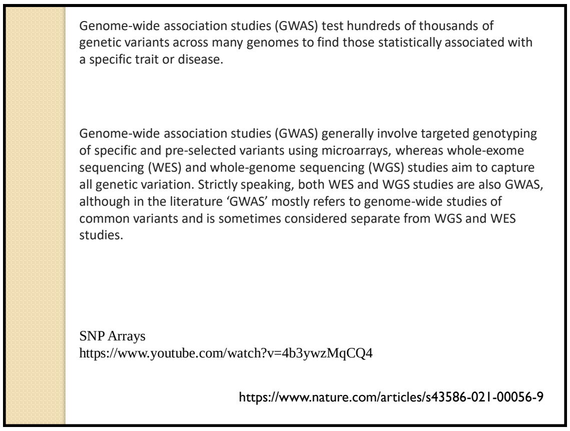
B&FG 3e
Page 1046



GENOME-WIDE ASSOCIATION STUDIES (GWAS)

A genome-wide association study (abbreviated GWAS) is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait. The method involves surveying the genomes of many people, looking for genomic variants that occur more frequently in those with a specific disease or trait compared to those without the disease or trait. Once such genomic variants are identified, they are typically used to search for nearby variants that contribute directly to the disease or trait.

https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies

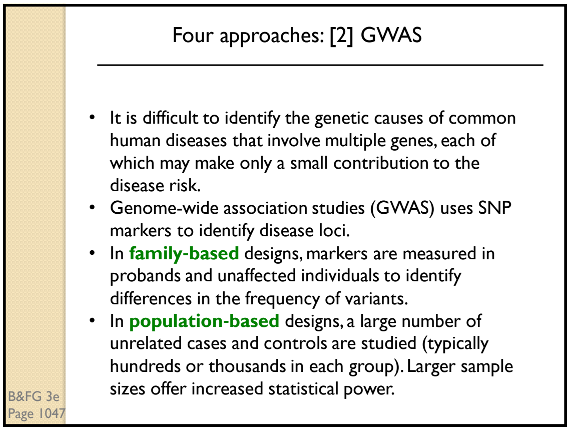


Genome-wide association studies (GWAS) test hundreds of thousands of genetic variants across many genomes to find those statistically associated with a specific trait or disease.

Genome-wide association studies (GWAS) generally involve targeted genotyping of specific and pre-selected variants using microarrays, whereas whole-exome sequencing (WES) and whole-genome sequencing (WGS) studies aim to capture all genetic variation. Strictly speaking, both WES and WGS studies are also GWAS, although in the literature 'GWAS' mostly refers to genome-wide studies of common variants and is sometimes considered separate from WGS and WES studies.

SNP Arrays
https://www.youtube.com/watch?v=4b3ywzMqCQ4

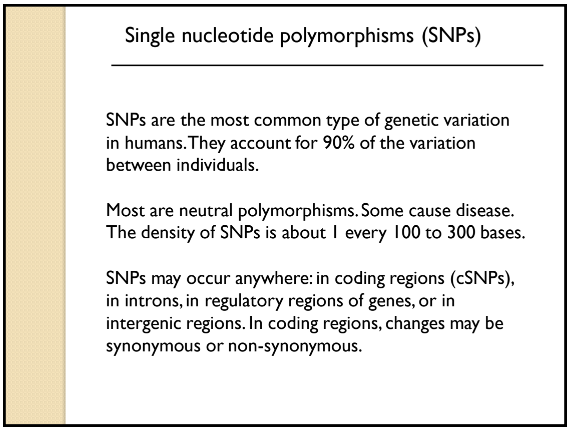
https://www.nature.com/articles/s43586-021-00056-9



Four approaches: [2] GWAS

- It is difficult to identify the genetic causes of common human diseases that involve multiple genes, each of which may make only a small contribution to the disease risk.
- Genome-wide association studies (GWAS) uses SNP markers to identify disease loci.
- In **family-based** designs, markers are measured in probands and unaffected individuals to identify differences in the frequency of variants.
- In **population-based** designs, a large number of unrelated cases and controls are studied (typically hundreds or thousands in each group). Larger sample sizes offer increased statistical power.

B&FG 3e
Page 1047

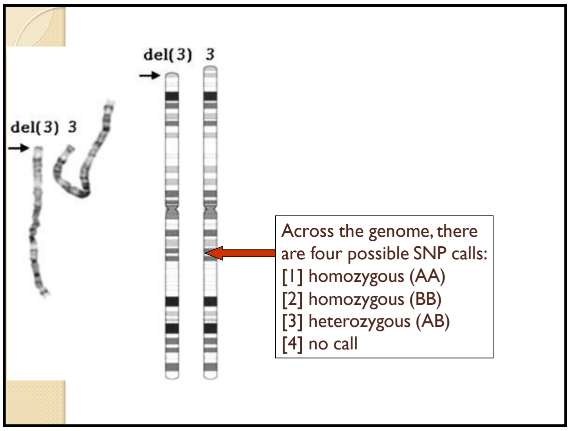


Single nucleotide polymorphisms (SNPs)

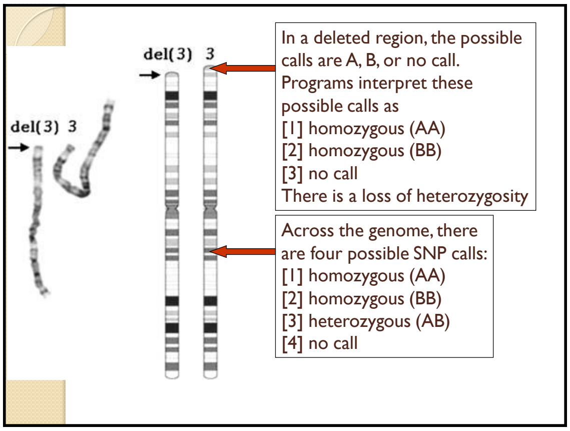
SNPs are the most common type of genetic variation in humans. They account for 90% of the variation between individuals.

Most are neutral polymorphisms. Some cause disease. The density of SNPs is about 1 every 100 to 300 bases.

SNPs may occur anywhere: in coding regions (cSNPs), in introns, in regulatory regions of genes, or in intergenic regions. In coding regions, changes may be synonymous or non-synonymous.



Across the genome, there are four possible SNP calls:
[1] homozygous (AA)
[2] homozygous (BB)
[3] heterozygous (AB)
[4] no call



In a deleted region, the possible calls are A, B, or no call. Programs interpret these possible calls as
[1] homozygous (AA)
[2] homozygous (BB)
[3] no call
There is a loss of heterozygosity

Across the genome, there are four possible SNP calls:
[1] homozygous (AA)
[2] homozygous (BB)
[3] heterozygous (AB)
[4] no call

SNPs and disease

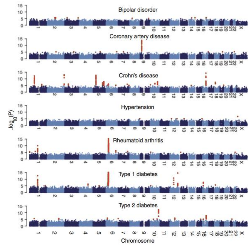
SNPs may be informative with respect to disease:

[1] Functional variation.A SNP associated with a nonsynonymous substitution in a coding region will change the amino acid sequence of a protein.

[2] Regulatory variation.A SNP in a noncoding region can influence gene expression.

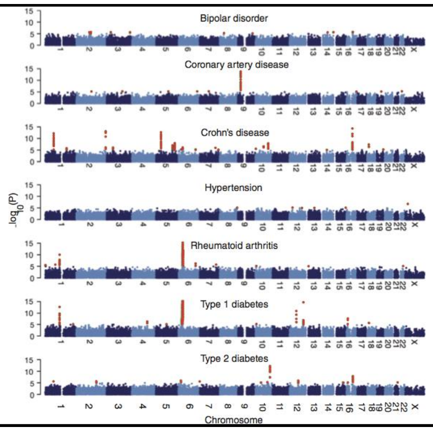
[3] Association. SNPs can be used in whole-genome association studies. SNP frequency is compared between affected and control populations.

Results of a genome-wide association study using 16,179 individuals to search for genes contributing to seven common familial disorders

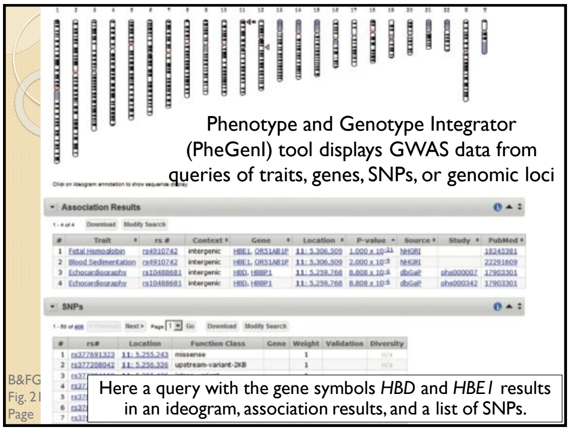


For each of seven diseases, the *y* axis shows the –log10 p value for SNPs that were positive for quality control criteria. The *x* axis shows the chromosomes. *p* values <1×10–5 are high-lighted in red. Panels are truncated at –log10(*p* value) = 15. Redrawn from Wellcome Trust Case Control Consortium (2007).

B&FG 3e
Fig. 21.17
Page 1049



B&FG 3e
Fig. 21.17
Page 1049



Phenotype and Genotype Integrator (PheGenI) tool displays GWAS data from queries of traits, genes, SNPs, or genomic loci

Click on ideogram annotation to show sequence display

Association Results

1 - 4 of 4 Download Modify Search

#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	PubMed
1	Fetal Hemoglobin	rs4910742	intergenic	HBE1, OR51AB1P	11: 5,306,509	1.000×10^{-21}	NHGRI		18245381
2	Blood Sedimentation	rs4910742	intergenic	HBE1, OR51AB1P	11: 5,306,509	2.000×10^{-8}	NHGRI		22291609
3	Echocardiography	rs10488681	intergenic	HBD, HBBP1	11: 5,259,768	8.808×10^{-6}	dbGaP	pha000007	17903301
4	Echocardiography	rs10488681	intergenic	HBD, HBBP1	11: 5,259,768	8.808×10^{-6}	dbGaP	pha000342	17903301

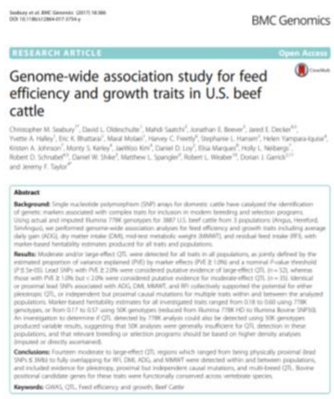
SNPs

1 - 60 of 406 < Previous Next > Page 1 Go Download Modify Search

#	rs#	Location	Function Class	Gene	Weight	Validation	Diversity
1	rs377691323	11: 5,255,243	missense		1		n/a
2	rs377208042	11: 5,256,326	upstream-variant-2KB		1		n/a
3	rs37						
4	rs37						
5	rs37						
6	rs37						
7	rs37						

B&FG
Fig. 21
Page

Here a query with the gene symbols *HBD* and *HBE1* results in an ideogram, association results, and a list of SNPs.



Seabury et al. BMC Genomics (2017) 18:386
DOI 10.1186/s12864-017-3754-y

BMC Genomics

RESEARCH ARTICLE

Open Access

Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle

Christopher M. Seabury[1*], David L. Oldeschulte[1], Mahdi Saatchi[2], Jonathan E. Beever[3], Jared E. Decker[4,5], Yvette A. Halley[1], Eric K. Bhattarai[1], Maral Molaei[1], Harvey C. Freetly[6], Stephanie L. Hansen[2], Helen Yampara-Iquise[4], Kristen A. Johnson[7], Monty S. Kerley[4], JaeWoo Kim[4], Daniel D. Loy[2], Elisa Marques[8], Holly L. Neibergs[7], Robert D. Schnabel[4,5], Daniel W. Shike[3], Matthew L. Spangler[9], Robert L. Weaber[10], Dorian J. Garrick[2,11] and Jeremy F. Taylor[4*]

Abstract

Background: Single nucleotide polymorphism (SNP) arrays for domestic cattle have catalyzed the identification of genetic markers associated with complex traits for inclusion in modern breeding and selection programs. Using actual and imputed Illumina 778K genotypes for 3887 U.S. beef cattle from 3 populations (Angus, Hereford, SimAngus), we performed genome-wide association analyses for feed efficiency and growth traits including average daily gain (ADG), dry matter intake (DMI), mid-test metabolic weight (MMWT), and residual feed intake (RFI), with marker-based heritability estimates produced for all traits and populations.

Results: Moderate and/or large-effect QTL were detected for all traits in all populations, as jointly defined by the estimated proportion of variance explained (PVE) by marker effects (PVE ≥ 1.0%) and a nominal *P*-value threshold (P ≤ 5e-05). Lead SNPs with PVE ≥ 2.0% were considered putative evidence of large-effect QTL (n = 52), whereas those with PVE ≥ 1.0% but < 2.0% were considered putative evidence for moderate-effect QTL (n = 35). Identical or proximal lead SNPs associated with ADG, DMI, MMWT, and RFI collectively supported the potential for either pleiotropic QTL, or independent but proximal causal mutations for multiple traits within and between the analyzed populations. Marker-based heritability estimates for all investigated traits ranged from 0.18 to 0.60 using 778K genotypes, or from 0.17 to 0.57 using 50K genotypes (reduced from Illumina 778K HD to Illumina Bovine SNP50). An investigation to determine if QTL detected by 778K analysis could also be detected using 50K genotypes produced variable results, suggesting that 50K analyses were generally insufficient for QTL detection in these populations, and that relevant breeding or selection programs should be based on higher density analyses (imputed or directly ascertained).

Conclusions: Fourteen moderate to large-effect QTL regions which ranged from being physically proximal (lead SNPs ≤ 3Mb) to fully overlapping for RFI, DMI, ADG, and MMWT were detected within and between populations, and included evidence for pleiotropy, proximal but independent causal mutations, and multi-breed QTL. Bovine positional candidate genes for these traits were functionally conserved across vertebrate species.

Keywords: GWAS, QTL, Feed efficiency and growth, Beef Cattle

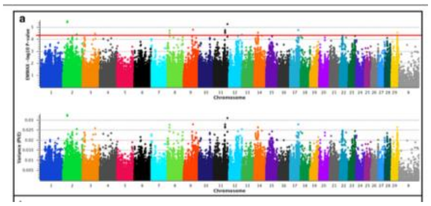
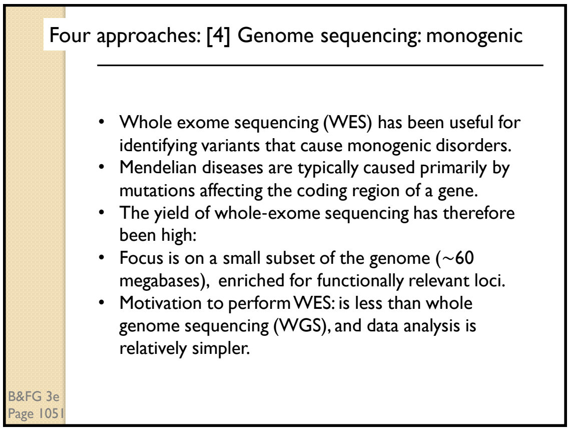


Fig. 1 Residual feed intake (RFI) QTL. The top pane of each composite panel reflects a Manhattan plot with EMMAX –log10 P-values for Illumina 778K markers, whereas the bottom pane depicts the estimated proportion of variance explained (PVE) by marker effects. Lead and supporting SNPs for moderate (1.0% < PVE < 2.0%) or large-effect QTL (PVE ≥ 2.0%) with P ≤ 5e-05 and MAF ≥ 0.01 are shown at or above the red line for U.S. Angus (a; n = 706)

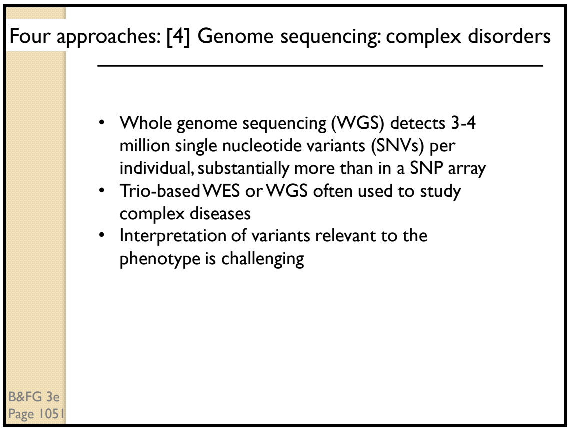
(RFI), with this trait often defined as the difference between an animal's observed and expected feed intake in relation to the animal's body weight and growth rate during a specified feeding period



Four approaches: [4] Genome sequencing: monogenic

- Whole exome sequencing (WES) has been useful for identifying variants that cause monogenic disorders.
- Mendelian diseases are typically caused primarily by mutations affecting the coding region of a gene.
- The yield of whole-exome sequencing has therefore been high:
- Focus is on a small subset of the genome (~60 megabases), enriched for functionally relevant loci.
- Motivation to perform WES: is less than whole genome sequencing (WGS), and data analysis is relatively simpler.

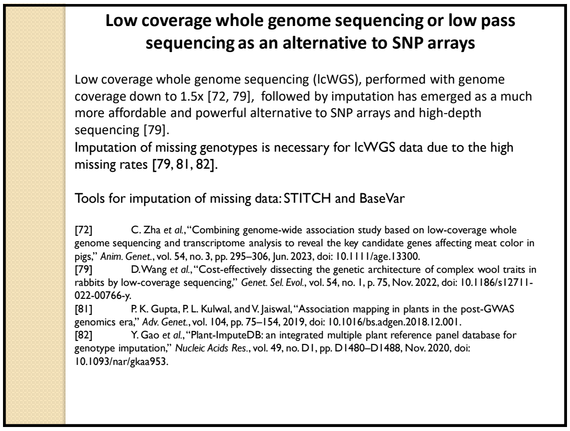
B&FG 3e
Page 1051



Four approaches: [4] Genome sequencing: complex disorders

- Whole genome sequencing (WGS) detects 3-4 million single nucleotide variants (SNVs) per individual, substantially more than in a SNP array
- Trio-based WES or WGS often used to study complex diseases
- Interpretation of variants relevant to the phenotype is challenging

B&FG 3e
Page 1051

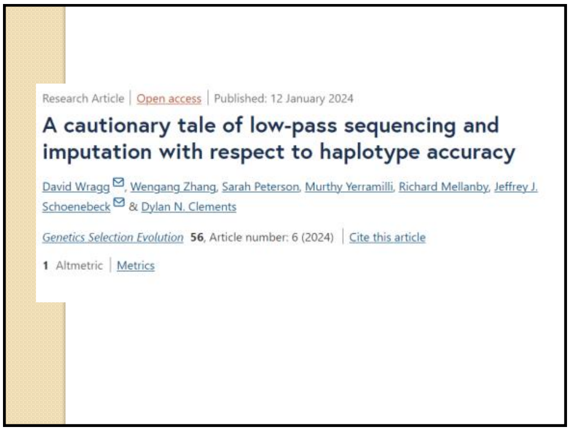


Low coverage whole genome sequencing or low pass sequencing as an alternative to SNP arrays

Low coverage whole genome sequencing (lcWGS), performed with genome coverage down to 1.5x [72, 79], followed by imputation has emerged as a much more affordable and powerful alternative to SNP arrays and high-depth sequencing [79].
Imputation of missing genotypes is necessary for lcWGS data due to the high missing rates [79, 81, 82].

Tools for imputation of missing data: STITCH and BaseVar

[72] C. Zha *et al.*, "Combining genome-wide association study based on low-coverage whole genome sequencing and transcriptome analysis to reveal the key candidate genes affecting meat color in pigs," *Anim. Genet.*, vol. 54, no. 3, pp. 295–306, Jun. 2023, doi: 10.1111/age.13300.
[79] D. Wang *et al.*, "Cost-effectively dissecting the genetic architecture of complex wool traits in rabbits by low-coverage sequencing," *Genet. Sel. Evol.*, vol. 54, no. 1, p. 75, Nov. 2022, doi: 10.1186/s12711-022-00766-y.
[81] P. K. Gupta, P. L. Kulwal, and V. Jaiswal, "Association mapping in plants in the post-GWAS genomics era," *Adv. Genet.*, vol. 104, pp. 75–154, 2019, doi: 10.1016/bs.adgen.2018.12.001.
[82] Y. Gao *et al.*, "Plant-ImputeDB: an integrated multiple plant reference panel database for genotype imputation," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1480–D1488, Nov. 2020, doi: 10.1093/nar/gkaa953.



Research Article | Open access | Published: 12 January 2024

A cautionary tale of low-pass sequencing and imputation with respect to haplotype accuracy

David Wragg, Wengang Zhang, Sarah Peterson, Murthy Yerramilli, Richard Mellanby, Jeffrey J. Schoenebeck & Dylan N. Clements

Genetics Selection Evolution **56**, Article number: 6 (2024) | Cite this article

1 Altmetric | Metrics