UNIVERSITY OF
HOHENHEIM

Genetic diversity: Genotyping and sequencing

3502-470 Plant Genetic Resources

Prof. Karl Schmid

SS 2025

Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

The Nature of Genetic Variation

The Nature of Genetic Variation

Types of genetic variation

- 1900’s: Visible polymorphisms
- 1930’s: Chromosomal polymorphisms
- 1940’s: Blood groups
- 1960’s: Protein polymorphisms
- 1980’s: DNA Sequencing
- 2000’s: Resequencing of genomes

What we will not teach in the module:

- Structure of DNA
- Functional elements of DNA (Genes, promotors, etc.)
- Genetic code, etc.
- Diversity of genetic markers

Types of DNA sequence variation

- Single nucleotide polymorphisms (SNPs)
- Insertion or Deletion variants (Indels)
- Structural genomic variants: Insertions or deletions from larger DNA seqments
- Variation in other repetitive elements such as minisatellites or gene families

Polymorphisms versus Markers

- Mutation: A genetic variant that was produced by a genetic process
- Polymorphism: A mutation that segregates in a
- Marker: A polymorphism that can be detected by a

Diversity of marker systems

- Co-dominant markers:
  - Single nucleotide polymorphisms (SNPs)
- Recent marker types:
  - Transposon 'counting' (many variants of this method)
  - Copy number variants (CNVs)
  - Presence/absence variants (PAVs)

SNP Genotyping

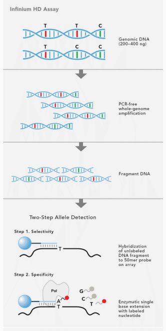
- Identification of SNPs by sequencing a small panel of individuals
- Design of a SNP array using a computer program
- Genotyping of many other individuals using the SNP array

SNP markers may range from 1 to 600,000 per array.

Genomic position is known: Multiple uses like genetic mapping, genetic diversity.

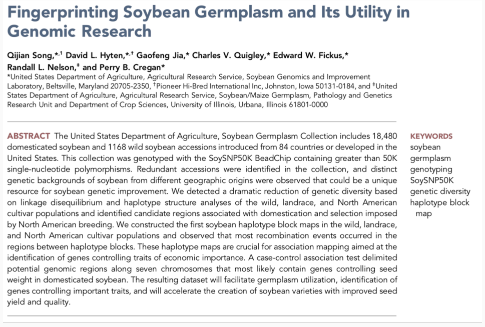
Key advantage: Low proportion of missing data!

SNP genotyping technology


Infinium HD Assay
T
T
C
T
C
C
Genomic DNA (200–400 ng)
PCR-free whole-genome amplification
Fragment DNA
Two-Step Allele Detection
Step 1. Selectivity
T
Hybridization of unlabeled DNA fragment to 50mer probe on array
Step 2. Specificity
Pol
G
C
A
T
T
Enzymatic single base extension with labeled nucleotide

The Infinium HD Assay protocol features single-tube sample preparation and whole-genome amplification without PCR or ligation steps, significantly reducing labor and sample-handling errors.

Infinium High Density Assay by Illumina



Fingerprinting Soybean Germplasm and Its Utility in Genomic Research

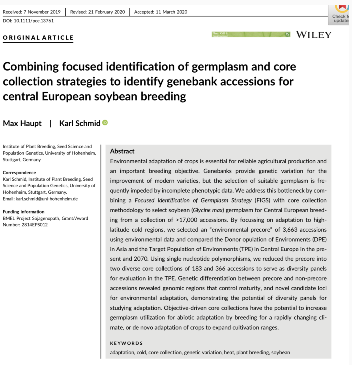
Qijian Song,*,[1] David L. Hyten,*,[†] Gaofeng Jia,* Charles V. Quigley,* Edward W. Fickus,*
Randall L. Nelson,[‡] and Perry B. Cregan*
*United States Department of Agriculture, Agricultural Research Service, Soybean Genomics and Improvement Laboratory, Beltsville, Maryland 20705-2350, [†]Pioneer Hi-Bred International Inc, Johnston, Iowa 50131-0184, and [‡]United States Department of Agriculture, Agricultural Research Service, Soybean/Maize Germplasm, Pathology and Genetics Research Unit and Department of Crop Sciences, University of Illinois, Urbana, Illinois 61801-0000

ABSTRACT The United States Department of Agriculture, Soybean Germplasm Collection includes 18,480 domesticated soybean and 1168 wild soybean accessions introduced from 84 countries or developed in the United States. This collection was genotyped with the SoySNP50K BeadChip containing greater than 50K single-nucleotide polymorphisms. Redundant accessions were identified in the collection, and distinct genetic backgrounds of soybean from different geographic origins were observed that could be a unique resource for soybean genetic improvement. We detected a dramatic reduction of genetic diversity based on linkage disequilibrium and haplotype structure analyses of the wild, landrace, and North American cultivar populations and identified candidate regions associated with domestication and selection imposed by North American breeding. We constructed the first soybean haplotype block maps in the wild, landrace, and North American cultivar populations and observed that most recombination events occurred in the regions between haplotype blocks. These haplotype maps are crucial for association mapping aimed at the identification of genes controlling traits of economic importance. A case-control association test delimited potential genomic regions along seven chromosomes that most likely contain genes controlling seed weight in domesticated soybean. The resulting dataset will facilitate germplasm utilization, identification of genes controlling important traits, and will accelerate the creation of soybean varieties with improved seed yield and quality.

KEYWORDS
soybean
germplasm
genotyping
SoySNP50K
genetic diversity
haplotype block map

G3·Genes | Genomes | Genetics

DOI: 10.1534/g3.115.019000 (2015)



Received: 7 November 2019 | Revised: 21 February 2020 | Accepted: 11 March 2020
DOI: 10.1111/pce.13761

ORIGINAL ARTICLE

WILEY

Combining focused identification of germplasm and core collection strategies to identify genebank accessions for central European soybean breeding

Max Haupt | Karl Schmid

Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

Correspondence
Karl Schmid, Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany.
Email: karl.schmid@uni-hohenheim.de

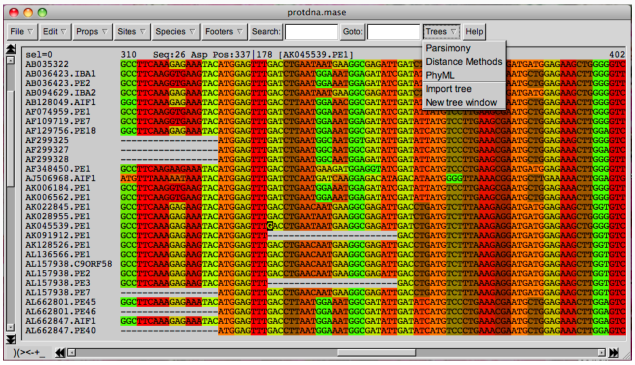
Funding information
BMEL Project Sojagenopath, Grant/Award Number: 2814EPS012

Abstract

Environmental adaptation of crops is essential for reliable agricultural production and an important breeding objective. Genebanks provide genetic variation for the improvement of modern varieties, but the selection of suitable germplasm is frequently impeded by incomplete phenotypic data. We address this bottleneck by combining a *Focused Identification of Germplasm Strategy* (FIGS) with core collection methodology to select soybean (*Glycine max*) germplasm for Central European breeding from a collection of >17,000 accessions. By focussing on adaptation to high-latitude cold regions, we selected an "environmental precore" of 3,663 accessions using environmental data and compared the Donor opulation of Environments (DPE) in Asia and the Target Population of Environments (TPE) in Central Europe in the present and 2070. Using single nucleotide polymorphisms, we reduced the precore into two diverse core collections of 183 and 366 accessions to serve as diversity panels for evaluation in the TPE. Genetic differentiation between precore and non-precore accessions revealed genomic regions that control maturity, and novel candidate loci for environmental adaptation, demonstrating the potential of diversity panels for studying adaptation. Objective-driven core collections have the potential to increase germplasm utilization for abiotic adaptation by breeding for a rapidly changing climate, or de novo adaptation of crops to expand cultivation ranges.

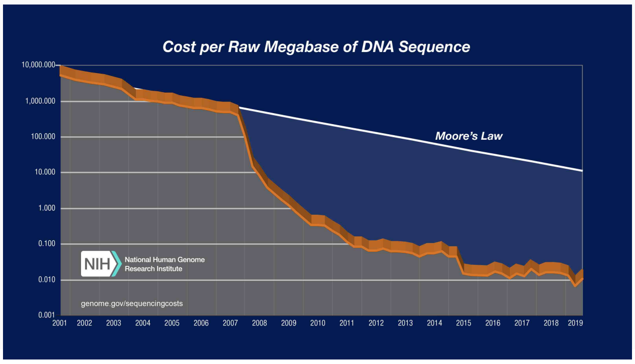
KEYWORDS
adaptation, cold, core collection, genetic variation, heat, plant breeding, soybean

Sequence variation: Example of a multiple sequence alignment

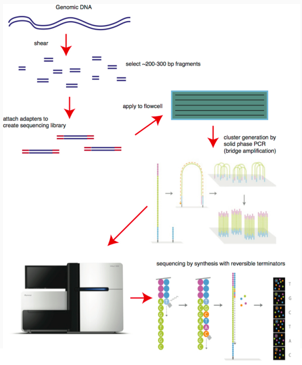


```
sel=0            310   Seq:26 Asp Pos:337|178 [AK045539.PE1]                                        402
AB035322         GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCT               GATGATGGAGAAGCTGGGGGTC
AB036423.IBA1    GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAT               AATGCTGGAGAAACTTGGGGTT
AB036423.PE2     GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAT               AATGCTGGAGAAACTTGGGGTT
AB094629.IBA2    GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCT               GATGATGGAGAAGCTGGGGGTC
AB128049.AIF1    GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAAACGGCGATATTGATAT               AATGCTGGAGAAACTTGGGGTC
AF074959.PE1     GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCTTGAAGCGAATGCTGGAGAAACTTGGGGTT
AF109719.PE7     GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCTTGAAGCGAATGCTGGAGAAACTTGGGGTT
AF129756.PE18    GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTC
AF299325         ------------------ATGGAGTTTGATCTGAATGGCAATGGTGATATTGATATTATGTCCCTGAAACGAATGCTGGAGAAGCTTGGGGTC
AF299327         ------------------ATGGAGTTTGATCTGAATGGCAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTC
AF299328         ------------------ATGGAGTTTGATCTGAATGGCAATGGAGATATCGATATTATGTCCTTGAAGCGAATGCTGGAGAAACTTGGGGTT
AF348450.PE1     GCCTTCAAGAAGAAATACATGGAGTTTGACCTGAATGAAGATGGAGGTATCGATATCATGTCCCTGAAGCGAATGATGGAGAAACTTGGGGTT
AJ506968.AIF1    ATGTTTAAAAATAAATACATGGAGTTTGATCTCAATGATCAAGGAGACATAGACATAATGGGGTTAAAACGGATGCTTGAAAAACTTGGAGTG
AK006184.PE1     GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCTTGAAGCGAATGCTGGAGAAACTTGGGGTT
AK006562.PE1     GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCTTGAAGCGAATGCTGGAGAAACTTGGGGTT
AK022845.PE1     GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AK028955.PE1     GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTGGGGGTC
AK045539.PE1     GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTGGGGGTC
AK091912.PE1     GCCTTCAAAGAGAAGTACATGGAGTTT------------------------GACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AK128526.PE1     GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AL136566.PE1     GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AL157938.C9ORF58 GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AL157938.PE2     GCCTTCAAAGAGAAGTACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AL157938.PE3     GCCTTCAAAGAGAAGTACATGGAGTTT------------------------GACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AL157938.PE7     ------------------ATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC
AL662801.PE45    GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTC
AL662801.PE46    ------------------ATGGAGTTTGACCTTAATGGAAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTC
AL662847.AIF1    GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTC
AL662847.PE40    ------------------ATGGAGTTTGACCTTAATGGAAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTC
```

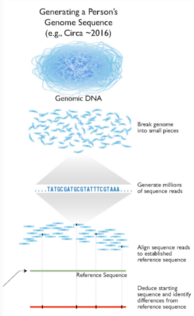
Dropping sequencing costs



Short read Illumina sequencing


Genomic DNA
shear
select ~200-300 bp fragments
attach adapters to create sequencing library
apply to flowcell
cluster generation by solid phase PCR (bridge amplification)
sequencing by synthesis with reversible terminators
T
G
C
T
A
C

Standardized short read sequencing workflows


Generating a Person's Genome Sequence (e.g., Circa ~2016)
Genomic DNA
Break genome into small pieces
....TATGCGATGCGTATTTCGTAAA....
Generate millions of sequence reads
Align sequence reads to established reference sequence
Reference Sequence
Deduce starting sequence and identify differences from reference sequence

National Human Genome Research Institute

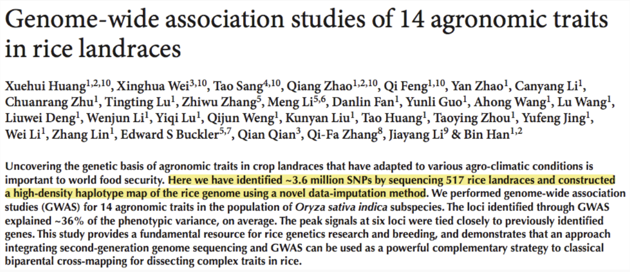
Newer technologies

Long read single molecule sequencing:

kilobases to megabase long sequence reads instead of hundreds of basepairs.

- PacBio
- Oxford Nanopore Minion

State of the art: Resequencing of complete genomes

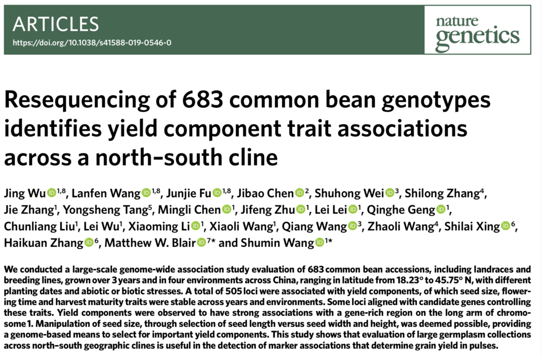


Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

Huang et al., Nature Genetics 2011



ARTICLES

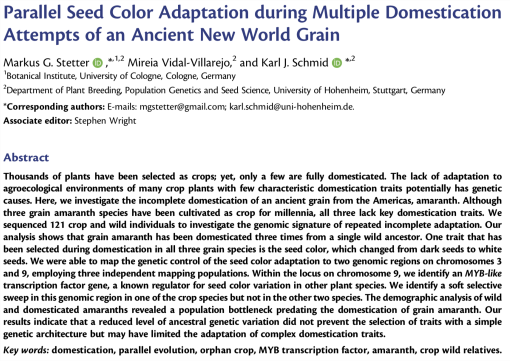
https://doi.org/10.1038/s41588-019-0546-0

nature genetics

Resequencing of 683 common bean genotypes identifies yield component trait associations across a north–south cline

Jing Wu[1,8], Lanfen Wang[1,8], Junjie Fu[1,8], Jibao Chen[2], Shuhong Wei[3], Shilong Zhang[4], Jie Zhang[1], Yongsheng Tang[5], Mingli Chen[1], Jifeng Zhu[1], Lei Lei[1], Qinghe Geng[1], Chunliang Liu[1], Lei Wu[1], Xiaoming Li[1], Xiaoli Wang[1], Qiang Wang[3], Zhaoli Wang[4], Shilai Xing[6], Haikuan Zhang[6], Matthew W. Blair[7*] and Shumin Wang[1*]

We conducted a large-scale genome-wide association study evaluation of 683 common bean accessions, including landraces and breeding lines, grown over 3 years and in four environments across China, ranging in latitude from 18.23° to 45.75° N, with different planting dates and abiotic or biotic stresses. A total of 505 loci were associated with yield components, of which seed size, flowering time and harvest maturity traits were stable across years and environments. Some loci aligned with candidate genes controlling these traits. Yield components were observed to have strong associations with a gene-rich region on the long arm of chromosome 1. Manipulation of seed size, through selection of seed length versus seed width and height, was deemed possible, providing a genome-based means to select for important yield components. This study shows that evaluation of large germplasm collections across north–south geographic clines is useful in the detection of marker associations that determine grain yield in pulses.



Parallel Seed Color Adaptation during Multiple Domestication Attempts of an Ancient New World Grain

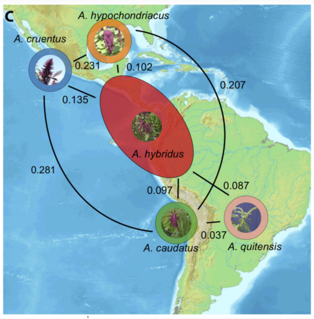
Markus G. Stetter ,*[1,2] Mireia Vidal-Villarejo,[2] and Karl J. Schmid *[2]
[1]Botanical Institute, University of Cologne, Cologne, Germany
[2]Department of Plant Breeding, Population Genetics and Seed Science, University of Hohenheim, Stuttgart, Germany
*Corresponding authors: E-mails: mgstetter@gmail.com; karl.schmid@uni-hohenheim.de.
Associate editor: Stephen Wright

Abstract

Thousands of plants have been selected as crops; yet, only a few are fully domesticated. The lack of adaptation to agroecological environments of many crop plants with few characteristic domestication traits potentially has genetic causes. Here, we investigate the incomplete domestication of an ancient grain from the Americas, amaranth. Although three grain amaranth species have been cultivated as crop for millennia, all three lack key domestication traits. We sequenced 121 crop and wild individuals to investigate the genomic signature of repeated incomplete adaptation. Our analysis shows that grain amaranth has been domesticated three times from a single wild ancestor. One trait that has been selected during domestication in all three grain species is the seed color, which changed from dark seeds to white seeds. We were able to map the genetic control of the seed color adaptation to two genomic regions on chromosomes 3 and 9, employing three independent mapping populations. Within the locus on chromosome 9, we identify an *MYB-like* transcription factor gene, a known regulator for seed color variation in other plant species. We identify a soft selective sweep in this genomic region in one of the crop species but not in the other two species. The demographic analysis of wild and domesticated amaranths revealed a population bottleneck predating the domestication of grain amaranth. Our results indicate that a reduced level of ancestral genetic variation did not prevent the selection of traits with a simple genetic architecture but may have limited the adaptation of complex domestication traits.

Key words: **domestication, parallel evolution, orphan crop, MYB transcription factor, amaranth, crop wild relatives.**

Multiple domestication of grain amaranths



Key technologies for analysing genetic variation

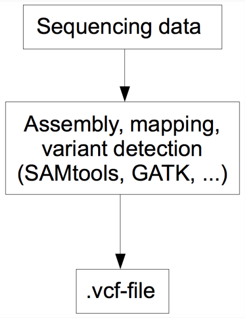
Genotyping: with SNP arrays

- Advantage: Only defined markers are used
- Disadvantage: Ascertainment bias

Sequencing: with “Next Generation Sequencing”

- Advantage: Complete genetic variation investigated
- Disadvantage: Missing data, sequence assembly

Bioinformatic analysis of sequencing data

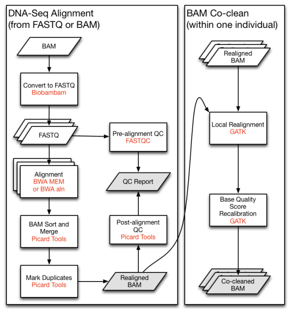


Sequencing data

Assembly, mapping, variant detection (SAMtools, GATK, ...)

.vcf-file

Bioinformatic analysis of sequencing data



vcf file format

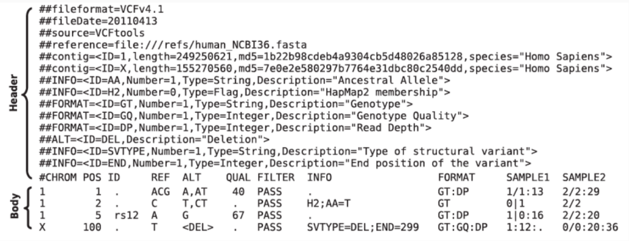
- **V**ariant **c**all **f**ormat
- Text file for storing variations like SNPs, indels or larger structural variants
- Actual version: VCF format v4.2
- Format description: `https://samtools.github.io/hts-specs/VCFv4.2.pdf`
- VCF files consist of a header and a body

vcf file format: Meta information

Meta-information `(## followed by key=value)`

- File format, mandatory (`##fileformat=VCFv4.1`)
- Additional information(`##samtoolsVersion=0.1.18 (r982:295)`)
- INFO lines (`##INFO=<ID=DP,Number=1,Type=Integer,Description= "Raw read depth">`)
- FORMAT lines (`##FORMAT=<ID=GT,Number=1,Type=String,Description= "Genotype">`)
- FILTER lines (`##FILTER=<ID=q10,Description="Quality below 10">`)

Example of meta information


Header
Body

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1 .      ACG  A,AT     40 PASS    .                    GT:DP      1/1:13   2/2:29
1        2 .      C    T,CT      . PASS    H2;AA=T              GT         0|1      2/2
1        5 rs12   A    G        67 PASS    .                    GT:DP      1|0:16   2/2:20
X      100 .      T    <DEL>     . PASS    SVTYPE=DEL;END=299   GT:GQ:DP   1:12:.   0/0:20:36
```

vcf file format: Header line

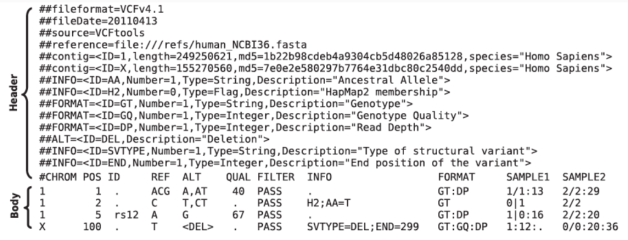
Mandatory columns: 8 fixed fields

- `CHROM` - chromosome id or number POS - reference position
- `ID` - unique identifier(s)
- `REF` - reference base(s)
- `ALT` - alternative base(s)
- `QUAL` - phred-scaled quality score for ALT
- `FILTER` - filters passed or not passed
- `INFO` - additional information, specified in meta-information

Optional columns:

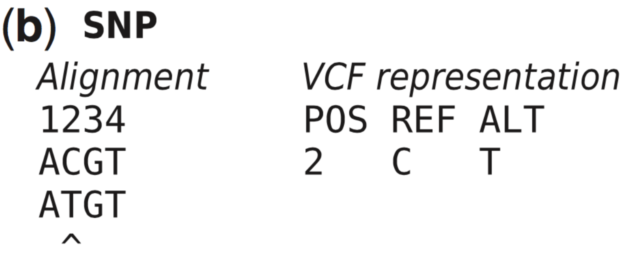
- `FORMAT` - information about genotype, read depth, etc.

vcf file format


Header
Body

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT    40  PASS    .                    GT:DP      1/1:13    2/2:29
1        2 .      C    T,CT    .   PASS    H2;AA=T              GT         0|1       2/2
1        5 rs12   A    G       67  PASS    .                    GT:DP      1|0:16    2/2:20
X      100 .      T    <DEL>   .   PASS    SVTYPE=DEL;END=299   GT:GQ:DP   1:12:.    0/0:20:36
```

Representation of polymorphisms



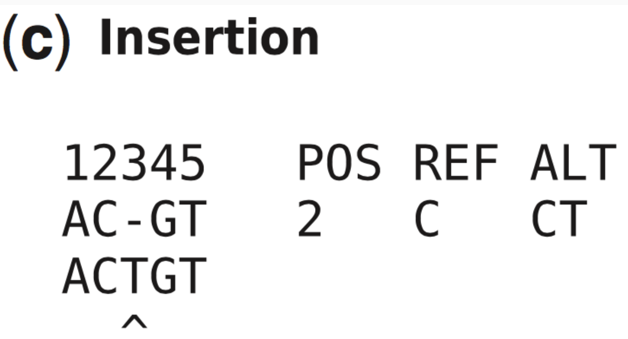
Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

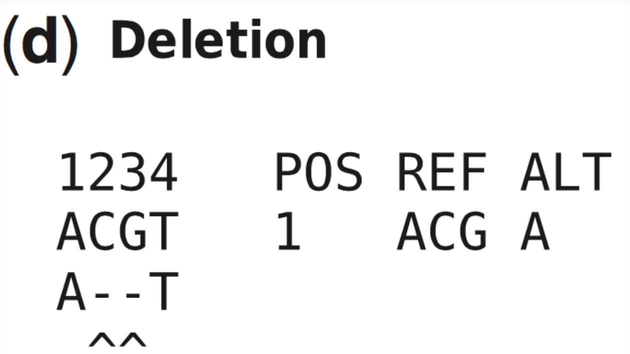
```
POS REF ALT
2   C   T
```

Representation of polymorphisms



```
12345   POS REF ALT
AC-GT   2   C   CT
ACTGT
   ^
```

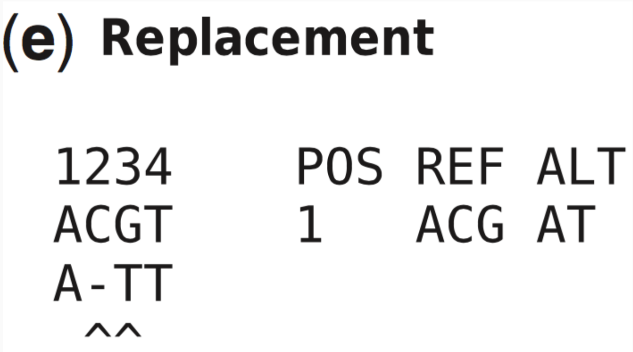
Representation of polymorphisms



(d) Deletion

```
1234   POS REF ALT
ACGT   1   ACG A
A--T
 ^^
```

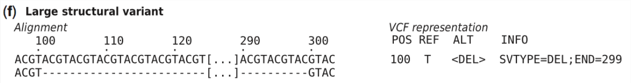
Representation of polymorphisms



(e) Replacement

```
1234    POS REF ALT
ACGT    1   ACG AT
A-TT
 ^^
```

Representation of polymorphisms



(f) Large structural variant

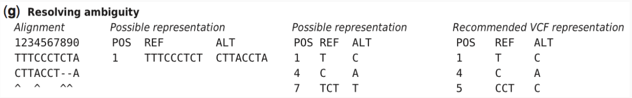
Alignment

```
   100       110       120       290       300
   .         .         .         .         .
ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT------------------------[...]----------GTAC
```

VCF representation

```
POS REF  ALT    INFO
100  T   <DEL>  SVTYPE=DEL;END=299
```

Representation of polymorphisms



(g) Resolving ambiguity

```
Alignment       Possible representation       Possible representation    Recommended VCF representation
1234567890      POS  REF        ALT           POS REF  ALT                POS   REF   ALT
TTTCCCTCTA      1    TTTCCCTCT  CTTACCTA      1   T    C                  1     T     C
CTTACCT--A                                    4   C    A                  4     C     A
^  ^   ^^                                     7   TCT  T                  5     CCT   C
```

Working with vcf files

- vcftools: Written in Perl, not so fast (`http://vcftools.github.io`)
- bcftools: Written in C very fast, fewer functions (`https://samtools.github.io/bcftools/`)

Further reading

- Metzker (2010) - Good, but a somewhat outdated review of sequencing technologies

References i

Metzker ML (2010) Sequencing technologies — the next generation. Nature Reviews Genetics 11(1):31–46, DOI 10.1038/nrg2626, URL http://www.nature.com/doifinder/10.1038/nrg2626