

# CircosVCF

**Sivan Gershanov**

**Pola Smirin-Yosef**

**Genomic Bioinformatics Laboratory**

**Department of Molecular Biology | Ariel University**

**Bioinformatics Workshops**

**09 Nov 2017**

**Tel Aviv, Israel**

# Outline

- Circos Visualization
- Variant Call Format (VCF) file
- CircosVCF
  - Introduction
  - Demo
  - Hands-on

**Circos Visualization**

Variant Call  
Format (VCF) file


CircosVCF

Demo

Hands-on

# Circos VISUALIZATION



**BECAUSE DINO DNA IS MILLIONS OF YEARS OLD, THERE ARE USUALLY GAPS IN THE CODE. OUR PALEOGENETICISTS USE DNA FROM RELATED SPECIES, SUCH AS BIRDS AND CROCODILES, TO FILL IN THE MISSING SEQUENCES.**

**THE TERRIFYING DINOSAUR CORN GENOME**

Amblin Entertainment and Legendary Pictures, the studios that produced Jurassic World, try to inject genome science into the movie. Unfortunately, since we don't quite know how to construct viable genomes of extinct species, much less grow the creatures themselves, we don't know whether the depiction of the science is right. Perhaps theirs is exactly what a genome lab would look like in a dino-building facility.

But, we can get fewer things wrong. In the [Creation Lab](#) companion website, a Circos image is used to illustrate a triceratops genome.

Unfortunately, this is an image of the B73 Maize reference genome (B73 RefGen\_v1), as published in Nature's [The B73 Maize Genome: Complexity, Diversity, and Dynamics](#).

Schnable PS Ware D Fulton RS *et al.* (2009) [The B73 maize genome: complexity, diversity, and dynamics](#) *Science* **326** (5956) 1112-1115

## WHAT IS CIRCOS?

### CIRCULAR VISUALIZATION

Circos is a software package for [visualizing data and information](#). It visualizes data in a [circular layout](#) — this makes Circos ideal for exploring relationships between objects or positions. There are [other reasons](#) why a circular layout is advantageous, not the least being the fact that it is attractive.

Circos is ideal for creating publication-quality infographics and illustrations with a high [data-to-ink ratio](#), richly layered data and pleasant symmetries. You have fine control each element in the figure to tailor its focus points and detail to your audience.

(Krzwinski et al., 2009) "Circos: an information aesthetic for comparative genomics". *Genome Research*, 19(9), 1639–45.

<http://circos.ca/>

# Circular Visualization

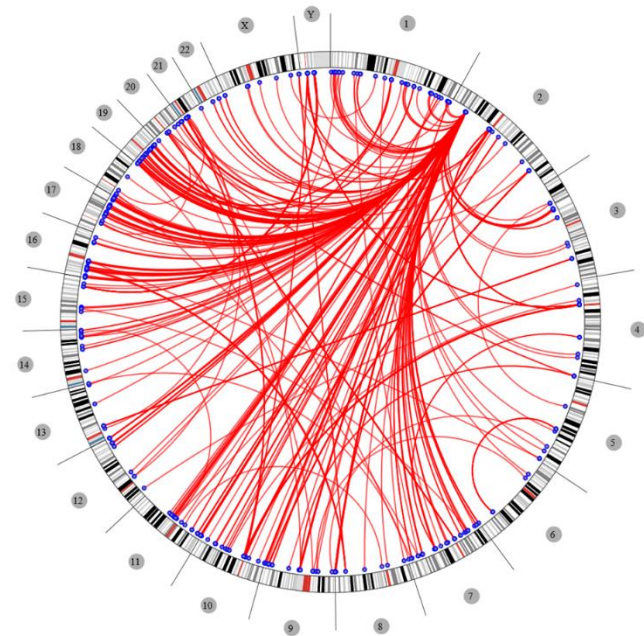
- Layering different data sets to create highly infographics with texture and visual appeal
- Integrate data for one or multiple samples to emphasize particularities, similarities or differences, in a single graphical representation
- Circos is capable of displaying data as:
  - Scatter line
  - Histogram plots
  - Heat maps
  - Tiles connectors and text

# 1. Genomic data

- Circular layout that facilitate sequenced genomic information
- Each chromosome is a segment around the circle

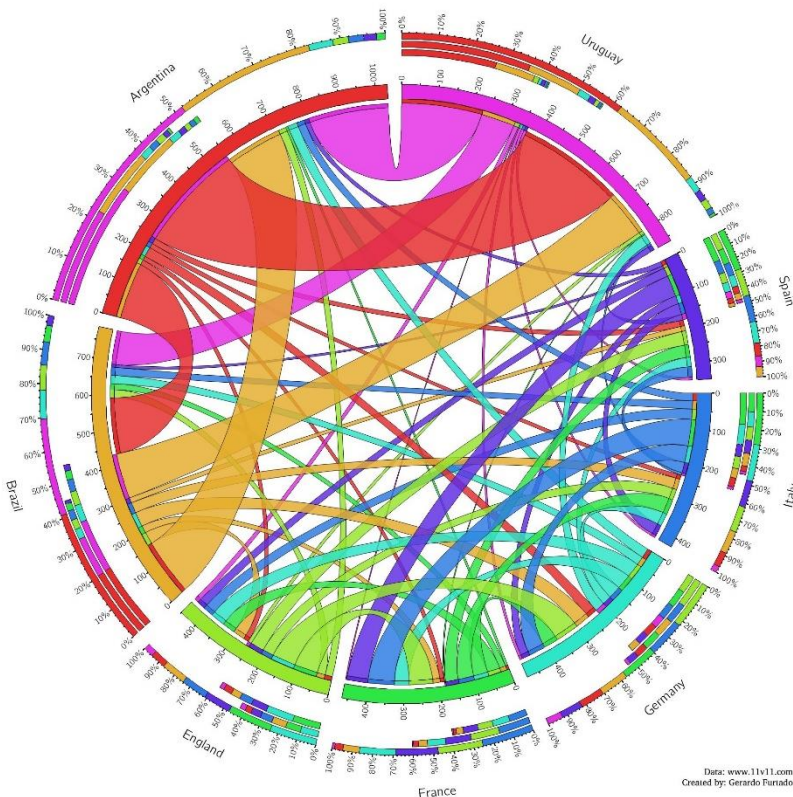
Features that have chromosomes and positions:

- Structural variants
- Repetitive elements
- Homology
- Evolutionary conservation scores
- SNPs
- Genes
- Differential gene expression
- DHS peaks
- Copy number profiles
- ChIP-seq peaks
- CpG islands



## 2. Chord diagram

- Inter-relationships between data in a matrix on one or more scales
- The size of a segment is the sum across a row or column in that adjacency matrix

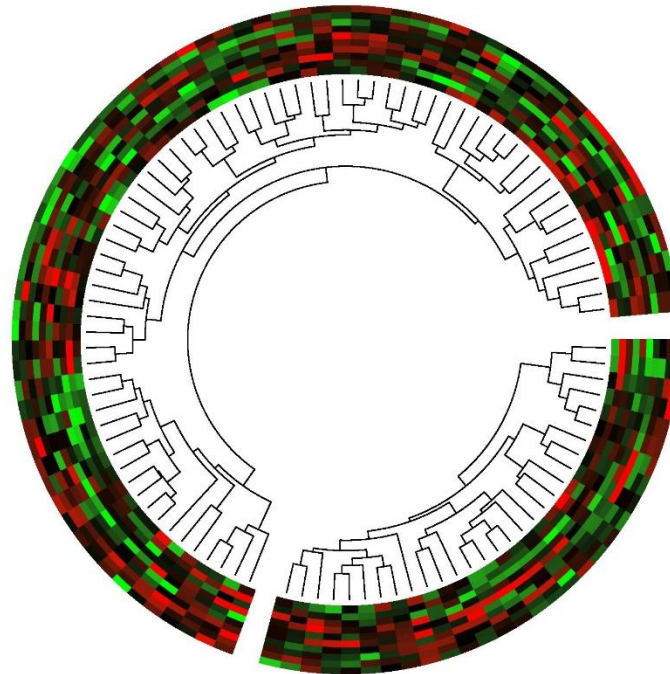


### ALL THE CHAMPIONS, ALL THE GOALS

All the goals scored and conceded by all the World Cup winners, in all matches held between them. Ribbons in the same colour of team mean goals scored, in different colours mean goals conceded.

# 3. Phylogeny

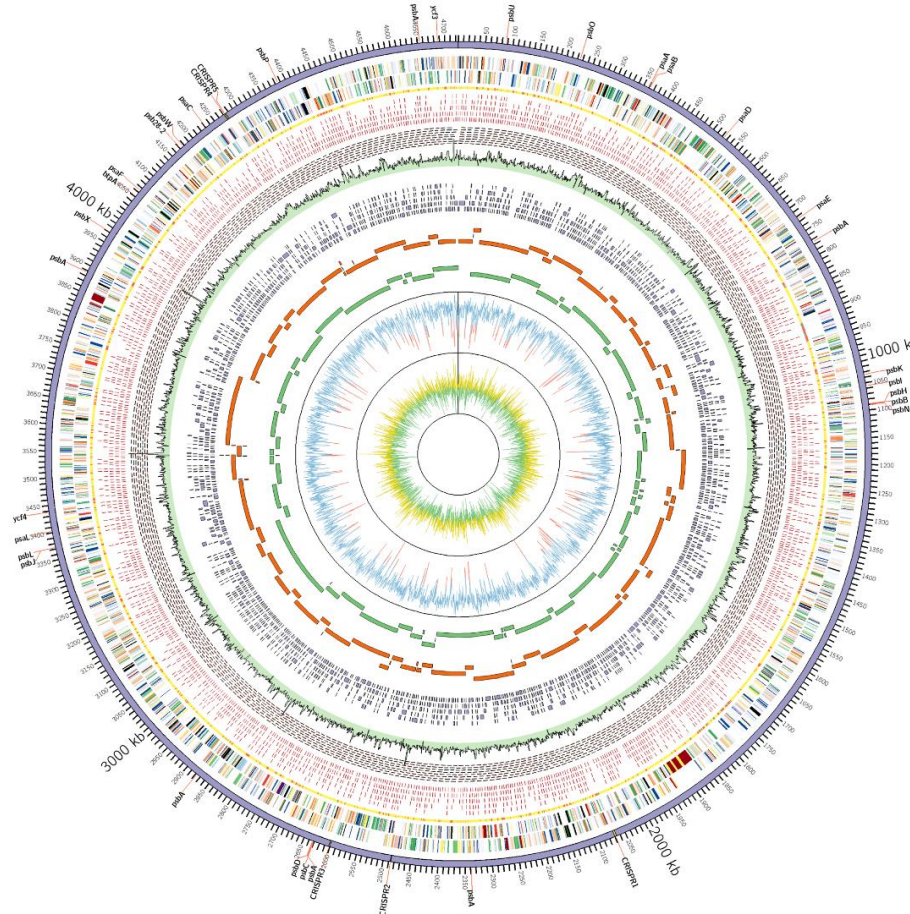
Distinct from both chromosomal coordinates and from chord diagrams (increased space for the leaf nodes)



A circos plot containing two phylogenetic trees with associated heatmaps



# Complicated Circos Plot



## Circular representation of the *Gloeobacter kilaueensis* JS1' genome

From inside out: GC skew (Yellow>0, Green<0), GC percent (Blue>50%, Red<50%), Newbler scaffold contigs, Celera contigs, Velvet contigs (Illumina reads only), read coverage (Combined 454 and Illumina reads sampled for 1,000 bp window. Highest coverage is 368 $\times$ ), minimal tiling clone pairs (shown in red), recruited reads from metagenome, taxonomic rank of top BLAST hit (yellow=Cyanobacteria, Red=others, Grey=no BLAST hit), coding regions in minus and plus strands (colored by COG functional categories). CRISPR repeat regions are highlighted in yellow in the outermost circle. Locations of genes involved in photosystems are labeled in the outermost circle

# Why to use Circos?

- Interesting data patterns
- Impressive figures for publications

# Visualization tools

- Circos – command line based tool -writing a long configuration file, challenging for users with no computing experience
- Circoletto
- J-Circos
- Circa – not free
- R based
  - RCircos
  - OmicCircos
  - Circlize
  - Ggbio
  - CIRCUS

<https://omictools.com/circos-plot-generation-category>

# Integrative Genomics Viewer (IGV)



Dark blue = heterozygous, Cyan = homozygous variant, Grey = reference.  
Filtered entries are transparent.

Circos Visualization

**Variant Call  
Format (VCF) file**

CircosVCF

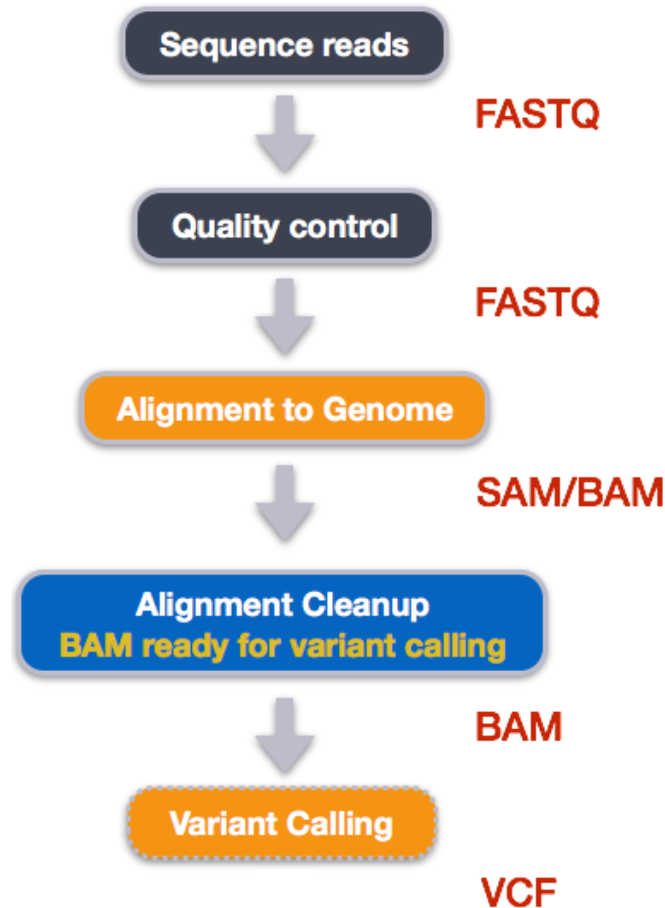
Demo

Hands-on

# Variant Call Format (VCF) file



# Sequencing data workflow



# Variant Callers

- Identification of variants from sequence data
- When aligned read differ from the reference genome it is written to a VCF file
- Over 40 open-source tools

GATK

Platypus

FreeBayes

VarScan

Samtools

LoFreq

Variant tools

VarDict

# Variant Call Format (VCF) file

- Text format
- Storing genotyping information
  - SNP
  - Indel
  - Structural variation
- Contain variants position
- The number of reads
- Overall quality



# Variant Call Format (VCF) file

Each line in the VCF file represents a single variant

Various properties of that variant represented in the columns

## Example

**VCF header**

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##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 (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##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">
```

**Mandatory header lines**

**Optional header lines** (meta-data about the annotations in the VCF body)

**Body**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

**Large SV**

**Insertion**

**Other event**

**Phased data** (G and C above are on the same chromosome)

# Fixed fields

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50
```

Field	Meaning
CHROM	Chromosome ID from the reference genome
POS	Reference position
ID	The rs number (dbSNP variant)
REF	Reference base (forward strand)
ALT	Alternate base observed in a sample/set of samples (forward strand)
QUAL	Probability that the ALT allele is <b>incorrectly</b> specified, expressed on the phred scale
FILTER	PASS if this position has passed all quality control filters
INFO	Additional information

# INFO field keys

- AA : ancestral allele
- AC : allele count in genotypes, for each ALT allele, in the same order as listed
- AF : allele frequency for each ALT allele in the same order as listed: use this when estimated from primary data, not called genotypes
- AN : total number of alleles in called genotypes
- BQ : RMS base quality at this position
- CIGAR : cigar string describing how to align an alternate allele to the reference allele
- DB : dbSNP membership
- DP : combined depth across samples, e.g. DP=154
- END : end position of the variant described in this record (for use with symbolic alleles)
- H2 : membership in hapmap2
- H3 : membership in hapmap3
- MQ : RMS mapping quality, e.g. MQ=52
- MQ0 : Number of MAPQ == 0 reads covering this record
- NS : Number of samples with data
- SB : strand bias at this position
- SOMATIC : indicates that the record is a somatic mutation, for cancer genomics
- VALIDATED : validated by follow-up experiment
- 1000G : membership in 1000 Genomes

# Genotype field

```
#CHROM POS ID REF ALT QUAL FILTER INFO
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017
```

```
FORMAT NA00001
GT:GQ:DP:HQ 0|0:48:1:51,51
GT:GQ:DP:HQ 0|0:49:3:58,50
```

## Field Meaning

GT	Genotype, encoded as allele values separated by either of / or  . 0/0 - the sample is homozygous reference 0/1 - the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles 1/1 - the sample is homozygous alternate
GQ	Genotype quality (Phred, probability that the genotype call (G/T) is correct)
DP	Filtered depth - number of filtered reads that support each of the reported alleles
HQ	Haplotype quality (Phred)

# Genotype field keys

Field	Number	Type	Description
AD	R	Integer	Read depth for each allele
ADF	R	Integer	Read depth for each allele on the forward strand
ADR	R	Integer	Read depth for each allele on the reverse strand
DP	1	Integer	Read depth
EC	A	Integer	Expected alternate allele counts
FT	1	String	Filter indicating if this genotype was “called”
GL	G	Float	Genotype likelihoods
GP	G	Float	Genotype posterior probabilities
GQ	1	Integer	Conditional genotype quality
GT	1	String	Genotype
HQ	2	Integer	Haplotype quality
MQ	1	Integer	RMS mapping quality
PL	G	Integer	Phred-scaled genotype likelihoods rounded to the closest integer
PQ	1	Integer	Phasing quality
PS	1	Integer	Phase set

Table 2: Reserved genotype fields

# VCF documentation

## The Variant Call Format (VCF) Version 4.2 Specification

24 May 2017

The master version of this document can be found at <https://github.com/samtools/hts-specs>.  
This printing is version 084587e from that repository, last modified on the date shown above.

### 1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

#### 1.1 An example

```
##fileformat=VCFv4.2
##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">
```

Circos Visualization

Variant Call  
Format (VCF) file

**CircosVCF**

Demo

Hands-on

# CircosVCF<sup>Beta</sup>

[legolas.ariel.ac.il/~tools/CircosVCF/](http://legolas.ariel.ac.il/~tools/CircosVCF/)

LET`S START

# CircosVCF

- Interactive user-friendly web service
- Create beautiful circos plots without writing a single line of code
- Whole genome variation information
- Comparison of variation distribution in multiple genomes
- Any organism and any reference genome
- No limit to the number of tracks or datasets you can plot
- Input: VCF files



# Ring types

- 1. Genotype:** Draw line in each SNP location
  - Color based on GT (0/0, 0/1, 1/1)
  - Recommended when plotting relatively small number of SNPs or variations in a single chromosome
- 2. Density:** SNP amount within a defined genomic length
  - Darker color represents denser regions

# Condition Types

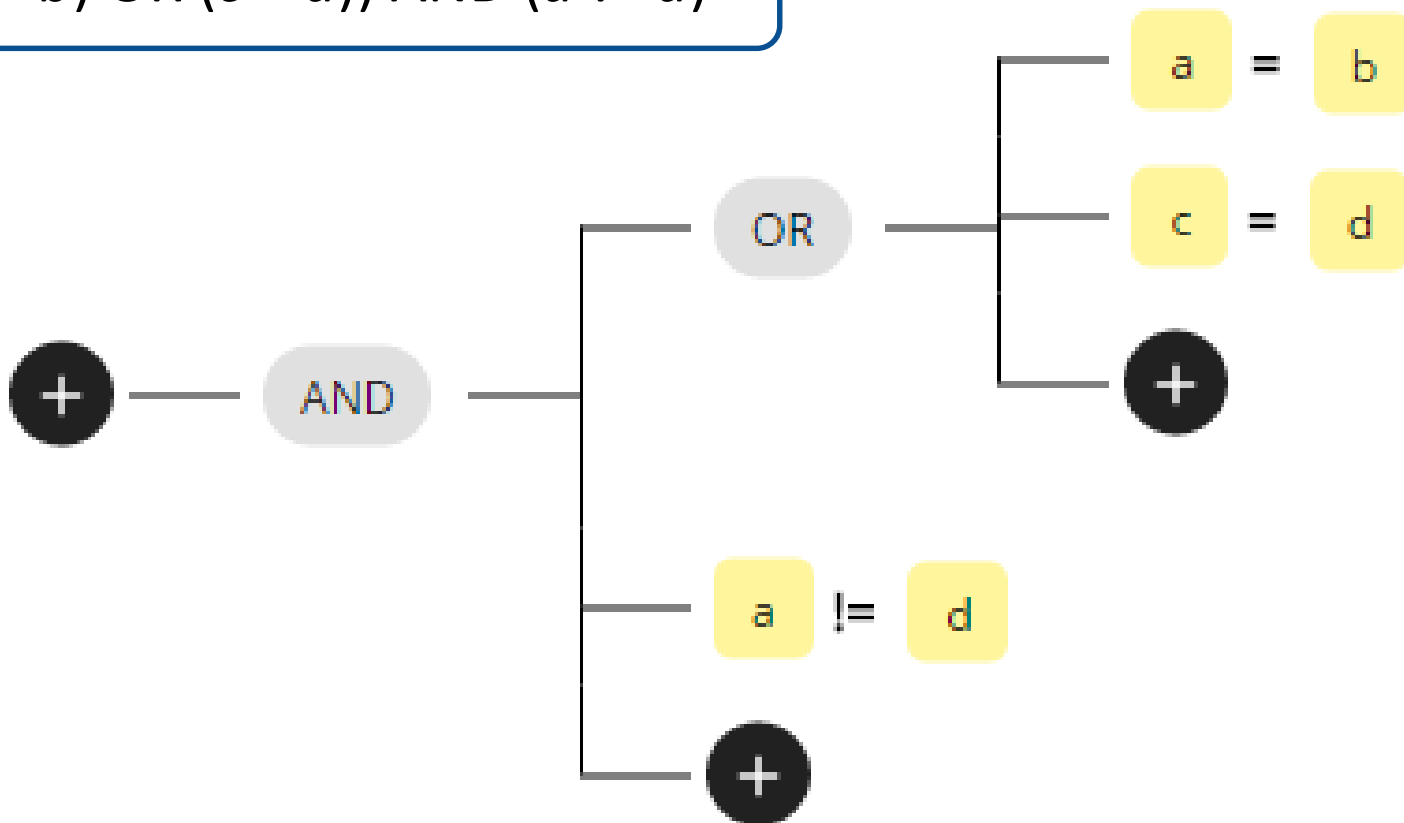
- AND Tree
- OR Tree
- Equal Columns
- Inequal Columns

# Comparison

- Has the possibility to **plot a subset of SNPs that would fulfill a specific set of conditions**

$$((a = b) \text{ OR } (c = d)) \text{ AND } (a \neq d)$$

plotting **all** SNPs where the genotypes of sample **a** are **equal** to sample **b**, **or** the genotypes of sample **c** are **equal** to sample **d**, as long as the genotypes of sample **a** are **different** from **d**.

$((a = b) \text{ OR } (c = d)) \text{ AND } (a \neq d)$ 

Circos Visualization

Variant Call  
Format (VCF) file

CircosVCF

**Demo**

Hands-on

DEMO

# Reviving the Wines of Ancient Israel

- Dr. Shivi Drori and Dr. Mali Salmon-Divon
  - Ariel University Wine Research Center
- Tracing the lost ancient varieties of wine grapes unique to the land of Israel
  - Remains were discovered in archaeological sites



# Reviving the Wines of Ancient Israel

- DNA sequencing and genetic analyses
  - Genetic map of the various strands
- Multi-VCF file of 9 *Vitis vinifera*
  - Sativa cultivars (domesticated) = 6
  - Sylvestris (wild) = 3



```

##contig=<ID=chrMT,length=773279>
##contig=<ID=chrPltd,length=160928>
##fileDate=20170822
##phasing=none
##reference=file:///data/genomes/vitis/VV12x.fa
##source=SelectVariants
##source=freeBayes v0.9.15-1-g076a2a2
##bcftools_viewVersion=1.3.1+htslib-1.3.1
##bcftools_viewCommand=view -s GEFEN16a,GEFEN1a,GEFEN21a,GEFEN23a,GEFEN2a.GEFEN9012.GEFEN9019.GEFEN9023.GEFEN9034 chr2.vcf
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sativa1_syl1 sativa2_syl2 syl3 sativa3 sativa4 sativa5 sativa6
chr2 462 . C T 51.45 . AB=0;ABP=0;AC=4;AF=1;AN=4;AO=5;CIGAR=1X;DP=12;DPB=12;DPRA=1.25;EPP=13.8677;EPPR=9.52472;GTI=4;LEN=1;MEAN
chr2 492 . T G 35.15 . AB=1;ABP=11.6962;AC=0;AF=0.0833333;AN=4;AO=4;CIGAR=1X;DP=29;DPB=29;DPRA=0.8;EPP=5.18177;EPPR=28.1125;GTI
chr2 499 . AT TG 39.15 . AB=1;ABP=11.6962;AC=0;AF=0.0833333;AN=4;AO=4;CIGAR=2X;DP=30;DPB=30.5;DPRA=0.769231;EPP=5.18177;EPPR=19.3
chr2 568 . ATAATA ATAATT,TTCATT 521.43 . AB=0.814815,1;ABP=26.2531,13.8677;AC=0,0;AF=0.0735294,0.0441176;AN=8;AO=22,5;CIGAR=5M1X,1X1M1X2M
chr2 579 . A T 526.67 . AB=0.814815;ABP=26.2531;AC=0;AF=0.0714286;AN=10;AO=22;CIGAR=1X;DP=186;DPB=186;DPRA=1.01887;EPP=4.58955;E
chr2 594 . T C 952.68 . AB=0.716981;ABP=24.6841;AC=2;AF=0.138889;AN=12;AO=38;CIGAR=1X;DP=270;DPB=270;DPRA=0.635023;EPP=3.0103;E
chr2 607 . GAAAAAAG GAAAAAAG 2053.21 . AB=0.412791;ABP=14.3727;AC=7;AF=0.27027;AN=14;AO=85;CIGAR=1M1I7M;DP=329;DPB=373.625;DPRA
chr2 615 . G C 5.93 . AB=0.666667;ABP=3.73412;AC=1;AF=0.0131579;AN=16;AO=2;CIGAR=1X;DP=389;DPB=389;DPRA=0.287565;EPP=3.0103;E
chr2 642 . AATC AATCTC 680.25 . AB=0.710526;ABP=17.6392;AC=2;AF=0.0657895;AN=16;AO=27;CIGAR=2M2I2M;DP=551;DPB=577.75;DPRA=0.488889;EPP=9
chr2 653 . C A 7246.36 . AB=0.492126;ABP=3.14709;AC=8;AF=0.368421;AN=16;AO=238;CIGAR=1X;DP=626;DPB=626;DPRA=1.44275;EPP=16.1851;E
chr2 665 . T C 2395.51 . AB=0.625954;ABP=21.0617;AC=2;AF=0.131579;AN=16;AO=82;CIGAR=1X;DP=659;DPB=659;DPRA=0.694697;EPP=6.82362;E
chr2 672 . G A 1392.13 . AB=0.588235;ABP=8.75832;AC=0;AF=0.0657895;AN=16;AO=51;CIGAR=1X;DP=681;DPB=681;DPRA=1.06103;EPP=3.05288;E
chr2 683 . A T 1369.28 . AB=0.592593;ABP=9.04217;AC=0;AF=0.0641026;AN=16;AO=48;CIGAR=1X;DP=665;DPB=665;DPRA=0.943151;EPP=4.6389;E
chr2 703 . C T 2127.73 . AB=0.462963;ABP=4.9405;AC=0;AF=0.102564;AN=16;AO=76;CIGAR=1X;DP=711;DPB=711;DPRA=1.13836;EPP=10.3247;E
chr2 722 . TCTC CCTT,CCTC 1067.46 . AB=0.409091,0.552632;ABP=7.74806,3.9246;AC=2,1;AF=0.0641026,0.0512821;AN=16;AO=27,21;CIGAR=1X2M1
chr2 739 . G A 124.02 . AB=0.352941;ABP=6.20364;AC=0;AF=0.0125;AN=18;AO=6;CIGAR=1X;DP=604;DPB=604;DPRA=1.12947;EPP=8.80089;EPPR=
chr2 758 . G T 5.58 . AB=0.181818;ABP=22.3561;AC=1;AF=0.025641;AN=18;AO=9;CIGAR=1X;DP=603;DPB=603;DPRA=1.28221;EPP=3.25157;E
chr2 782 . G A 0.55 . AB=0.25;ABP=7.35324;AC=1;AF=0.0128205;AN=18;AO=2;CIGAR=1X;DP=594;DPB=594;DPRA=0.518771;EPP=3.0103;EPPR=5
chr2 788 . C T 336.48 . AB=0.346939;ABP=12.9813;AC=2;AF=0.075;AN=18;AO=17;CIGAR=1X;DP=580;DPB=580;DPRA=0.522913;EPP=3.13803;EPP
chr2 836 . C A 968.11 . AB=0.795455;ABP=36.372;AC=2;AF=0.0769231;AN=16;AO=37;CIGAR=1X;DP=596;DPB=596;DPRA=0.598536;EPP=3.06899;E
GT:AO:DP:GQ:PL:QA:QR:RO 1/1:5:5:17:100,15,0:197:0:0 0/0:0:2:17:0,6,73:0:77:2 0/1:6:8:66:100,0,64:220:77:2
GT:AO:DP:GQ:PL:QA:QR:RO 0/0:0:6:29:0,18,100:0:227:6 0/0:0:2:17:0,6,73:0:77:2 0/0:0:9:38:0,27,100:0:330:9
GT:AO:DP:GQ:PL:QA:QR:RO 1/1:5:5:16:100,15,0:194:0:0 0/0:0:2:20:0,6,73:0:77:2 0/1:7:10:99:100,0,100:277:120:3
GT:AO:DP:GQ:PL:QA:QR:RO 0/0:0:5:50:0,15,100:0:197:5 0/1:2:2:0:76,6,0:80:0:0 0/0:0:10:65:0,30,100:0:388:10
GT:AO:DP:GQ:PL:QA:QR:RO 0/2:0,1:5:20:29,41,100,0,100,100:0:37:158:4 0/0:0,0:2:13:0,6,73,6,73,73:0,0:77:2 0/0:0,0:9:31:0,24,100,24,100,100:0,0:297:8
GT:AO:DP:GQ:PL:QA:QR:RO 0/0:0:6:56:0,18,100:0:228:6 0/0:0:1:41:0,3,40:0:40:1 0/0:0:16:86:0,48,100:0:586:16
GT:AO:DP:GQ:PL:QA:QR:RO 1/1:4,0:4:22:100,12,0,100,12,100:132,0:0:0 0/2:0,1:1:0:40,40,3,3,0:0,40:0:0 0/1:6,0:14:99:100,0,100,100,100,100:234,0:297:8
GT:AO:DP:GQ:PL:QA:QR:RO 1/1:4,0:4:18:100,12,0,100,12,100:144,0:0:0 2/2:0,1:1:14:40,40,40,3,3,0:0,40:0:0 0/1:6,0:14:99:100,0,100,100,100,100:237,0:301:8
GT:AO:DP:GQ:PL:QA:QR:RO 0/0:0:5:47:0,15,100:0:176:5 0/0:0:1:35:0,3,37:0:37:1 0/1:7:13:99:100,0,100:264:231:6

```

Input file: 6 gb



Circos Visualization

Variant Call  
Format (VCF) file

CircosVCF

**Demo**

Hands-on

CircosVCF<sup>Beta</sup>

**DEMO TIME**

[legolas.ariel.ac.il/~tools/CircosVCF/](http://legolas.ariel.ac.il/~tools/CircosVCF/)

LET`S START

Circos Visualization

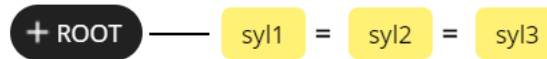
Variant Call  
Format (VCF) file

CircosVCF

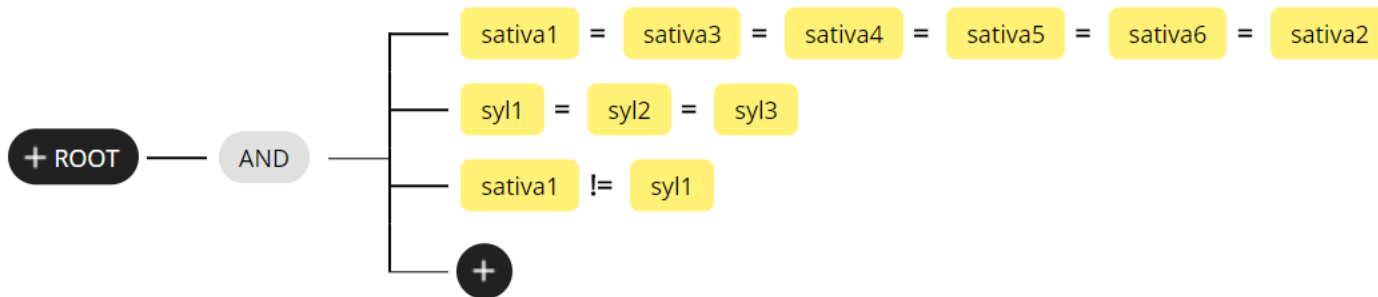
**Demo**

Hands-on

## Condition tree for density ring



## Condition tree for genotype ring



Circos Visualization

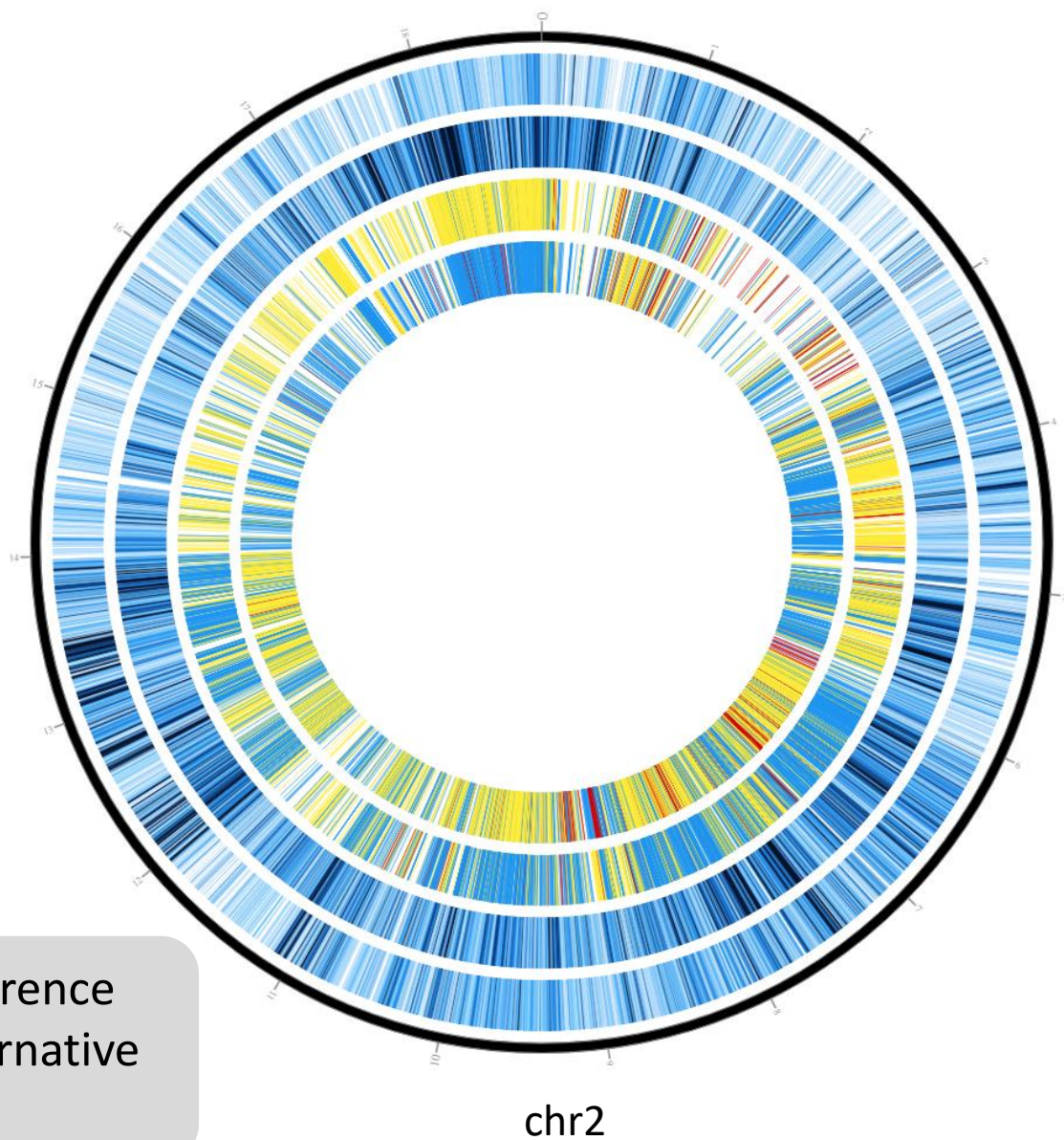
Variant Call  
Format (VCF) file

CircosVCF

Demo

Hands-on

Tracks/ Circles	
1	SatGroup – Density - SNPs distribution
2	SylGroup – Density - SNPs distribution
3	Condition tree - Sativa
4	Condition tree - Sylvestris



**Yellow** Homozygosity to the reference  
**Red** Homozygosity to the alternative  
**Blue** Heterozygosity

Circos Visualization

Variant Call  
Format (VCF) file

CircosVCF

Demo

**Hands-on**

# Hands-on

# Acknowledgment

**Genomic Bioinformatics Laboratory, Molecular Biology, Ariel University**

Dr. Mali Salmon-Divon

Doron Levi

Tamar Nehushtam

Tamar Harel

Tamar Segal

Deepak Karthik

Sne Morag



**THANK  
YOU**

**Sivan Gershanov**  
[sivang@ariel.ac.il](mailto:sivang@ariel.ac.il)

**Pola Smirin-Yosef**  
[polasy@ariel.ac.il](mailto:polasy@ariel.ac.il)