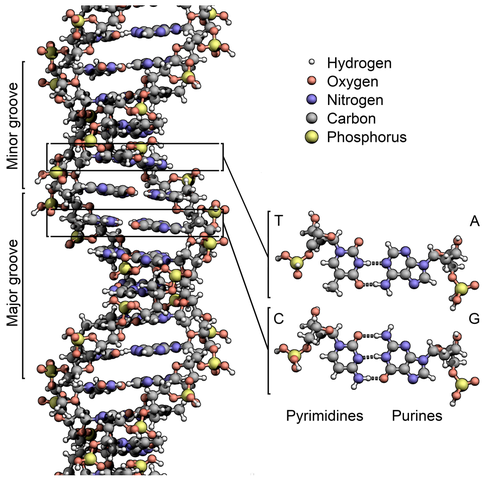
Biological concepts

- **closely related terms: gene, allele, mutation, variation, SNP**
- **gene** = distinct area on DNA, responsible for a trait of an individual (e.g. gene for coloration)
- **allele** = a version of a gene (coding for color A/color B)
- **allele frequency =** relative abundance of an allele in a population
- **mutation** = change in the DNA sequence, mostly describing the process of change, mutation causes variation
- **variation** = can refer to a single base in sequence or to even bigger units (structural variation, copy of a gene or copy of a chromosome), also referred to as polymorphism
- **SNP** = “Single Nucleotide Polymorphism”, variation in only one base/ nucleotide

1) What influences variation?



Wildtype **mutation**

T A A C C G A T T T A T

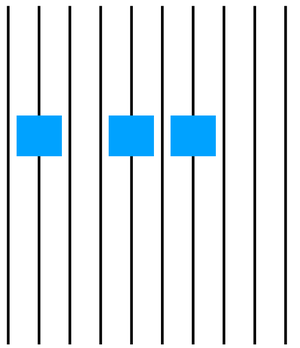
Allele

T A A C C A A T T T A T

genotype -> phenotype

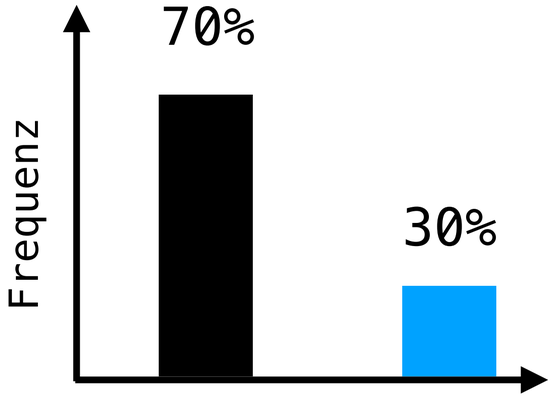
Monogenic trait

allele-frequency



G: 7/10 = 70%

A: 3/10 = 30%

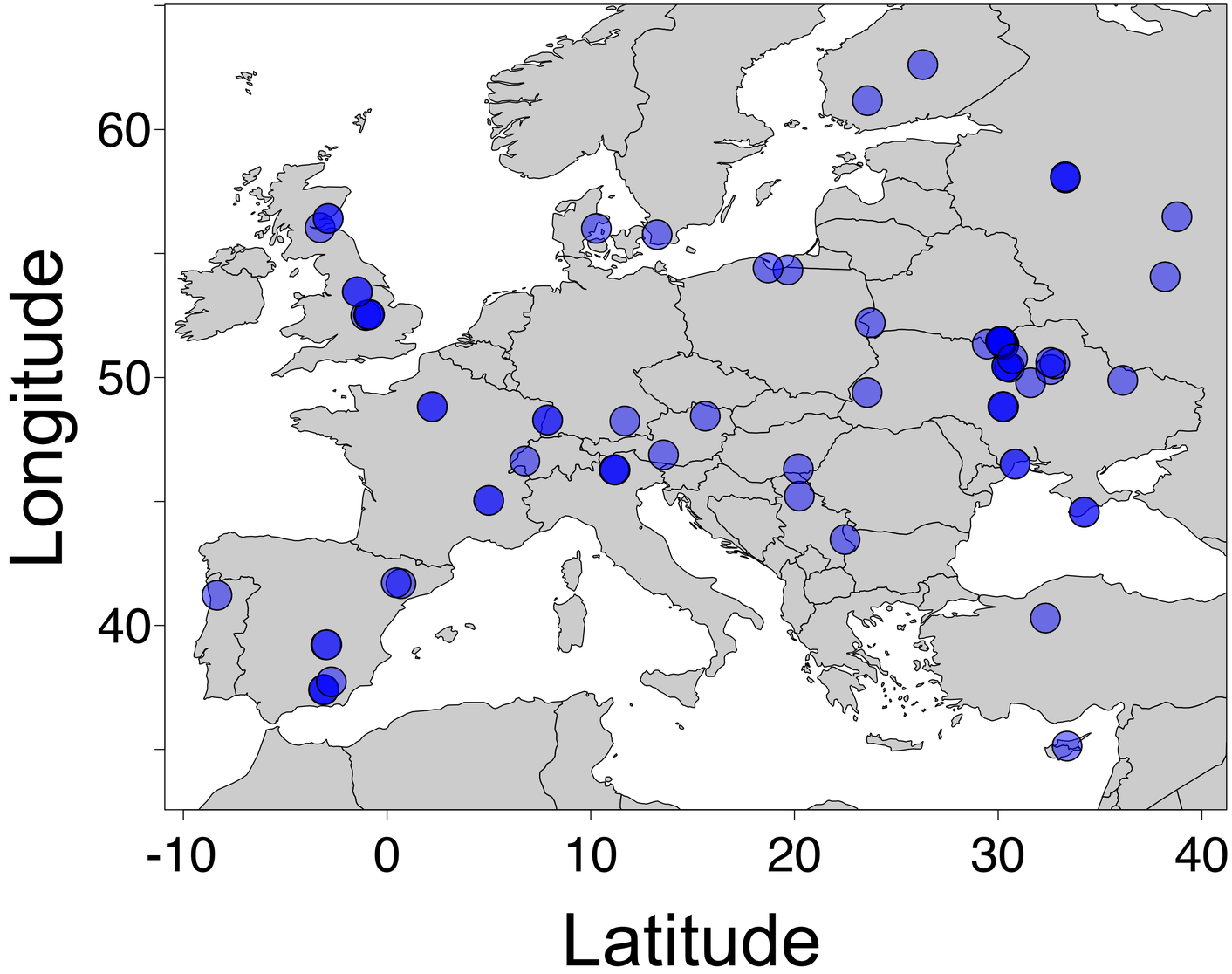


Phenotype: not only appearance, can also be behaviour or a trait (survival, performance in a certain habitat,…)

https://www.odaguides.us/cepaea-nemoralis.html

www.droseu.net

- **pan-european consortium**
- Since **2013**
- founded and organised by
  - Josefa Gonzalez
  - Martin Kapun
  - Thomas Flatt
- **>160 spatiotemporal Population-samples**
- **Pooled sequencing**



Kapun *et al.* (2020) MBE
Kapun *et al.* (2021) MBE

Reduce to relevant information

```
          Pos 1
           Pos 2
            Pos3
Indiv.1 ACTGCTAGACTAGCTAGCTAGCTACGATCGATC
Indiv.2 ACTGCTAGACTAGGTAGCTAGCTACGTTCGATC
Indiv.3 ACTGCTAGACTAGGTAGCTAGCTACGATCGATC
Indiv.4 ACTGCTAGACTAGCTAGCTAGCTACGTTCGATC
Indiv.5 ACTGCTAGACTAGCTAGCTAGCTACGATCGATC
Indiv.6 ACTGCTAGACTAGGTAGCTAGCTACGATCGATC
Indiv.7 ACTGCTAGACTAGCTAGCTAGCTACGTTCGATC
Indiv.8 ACTGCTAGACTAGGTAGCTAGCTACGATCGATC
```

```
ACTGCTAGACTAGCTAGCTAGCTACGATCGATC
ACTGCTAGACTAGGTAGCTAGCTACGTTCGATC
ACTGCTAGACTAGGTAGCTAGCTACGATCGATC
ACTGCTAGACTAGCTAGCTAGCTACGTTCGATC
ACTGCTAGACTAGCTAGCTAGCTACGATCGATC
ACTGCTAGACTAGGTAGCTAGCTACGATCGATC
```

Many more steps

Bioinformatic Pipeline

```
##contig=<ID=X>
##contig=<ID=Y>
##contig=<ID=4>
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact Gene_Name | Gene_ID
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this Format: 'Gene_Name |
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this Format: 'Gene_Name |
##bcftools_viewCommand=view -O b --threads 3; Date=Fri Jan 21:58:22 2021
##bcftools_viewCommand=view -O z /project/berglandlab/DEST/vcf/dest.PoolSeq.PoolSNP.001.50.10Nov2020.ann.vcf.gz; Date=Sat Jan 16 2021
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  FORMAT  AT_Mau_14_01    AT_Mau_14_02
2L      5437    .       A       T       .       .       GT:RD:AD:DP:FREQ        0/0:21:0:21:0   0/0:26:0:26:0
2L      5516    .       G       A       .       .       GT:RD:AD:DP:FREQ        0/0:26:0:26:0   0/0:31:0:31:0
2L      5750    .       A       T       .       .       GT:RD:AD:DP:FREQ        0/1:21:2:23:0.09        0/1:10:1:11:0.09
2L      5762    .       T       C       .       .       GT:RD:AD:DP:FREQ        0/1:15:4:19:0.21        0/1:7:6:13:0.46
2L      5776    .       C       G       .       .       GT:RD:AD:DP:FREQ        0/1:21:2:23:0.09        0/1:13:1:14:0.07
2L      5803    .       A       G       .       .       GT:RD:AD:DP:FREQ        0/1:25:2:27:0.07        0/0:13:0:13:0
2L      5813    .       G       T       .       .       GT:RD:AD:DP:FREQ        0/1:26:1:27:0.04        0/1:16:1:17:0.06
2L      5836    .       T       A       .       .       GT:RD:AD:DP:FREQ        0/0:20:0:20:0   0/1:11:2:13:0.15
2L      5845    .       C       T       .       .       GT:RD:AD:DP:FREQ        0/0:25:0:25:0   0/0:16:0:16:0
2L      6079    .       C       T       .       .       GT:RD:AD:DP:FREQ        0/1:27:2:29:0.07        0/1:22:1:23:0.04
2L      6166    .       T       C       .       .       GT:RD:AD:DP:FREQ        0/0:29:0:29:0   0/1:16:8:24:0.33
2L      6256    .       G       A       .       .       GT:RD:AD:DP:FREQ        0/0:26:0:26:0   0/1:27:2:29:0.07
2L      6321    .       C       T       .       .       GT:RD:AD:DP:FREQ        0/0:34:0:34:0   0/0:29:0:29:0
2L      6323    .       A       C       .       .       GT:RD:AD:DP:FREQ        0/0:34:0:34:0   0/0:31:0:31:0
2L      6353    .       C       T       .       .       GT:RD:AD:DP:FREQ        0/1:11:22:33:0.67       0/1:15:17:32:0.53
```

VCF = variant call format

Dummy Dataset

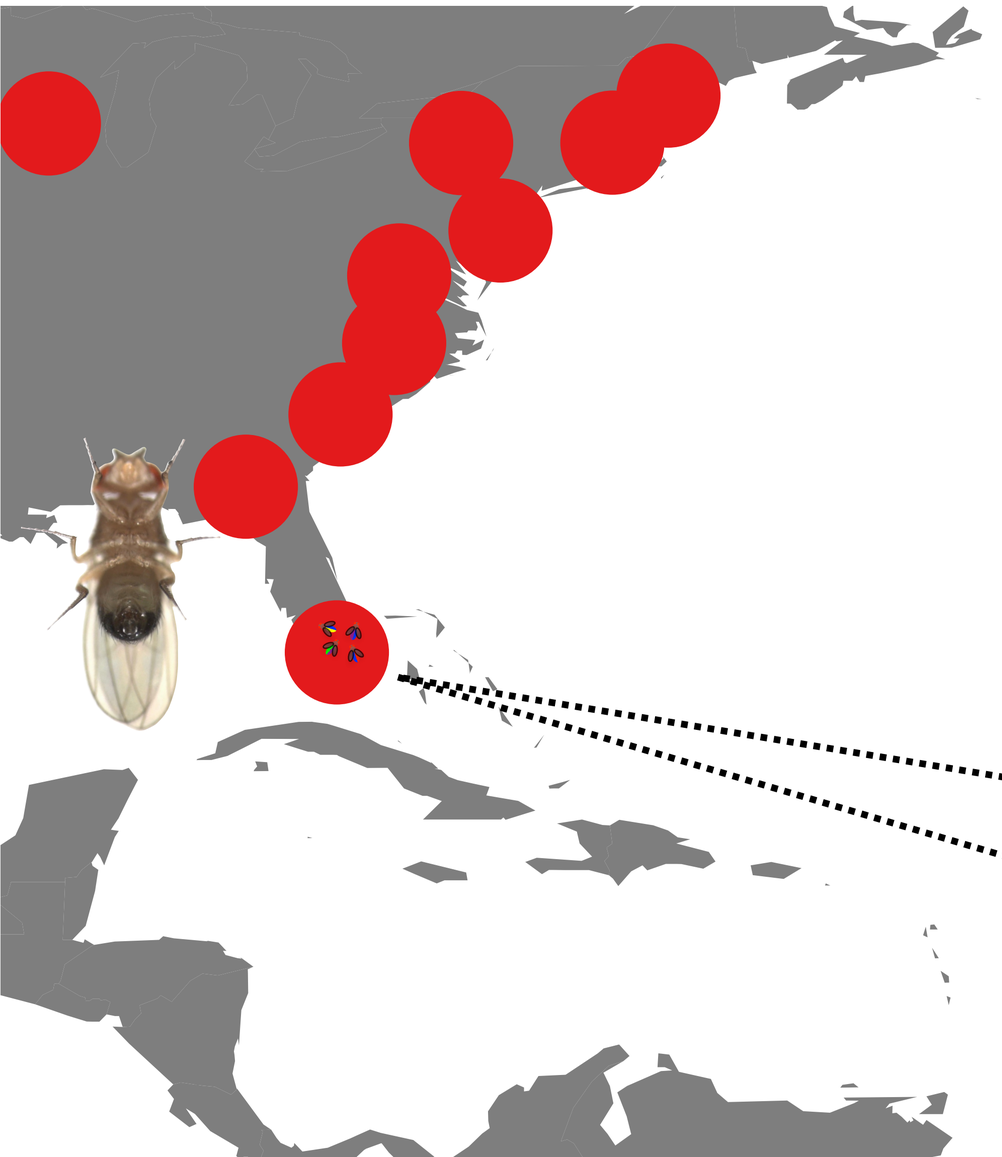
- **10 000 Positions**
- 9 Populationen
- 4 environmental variables

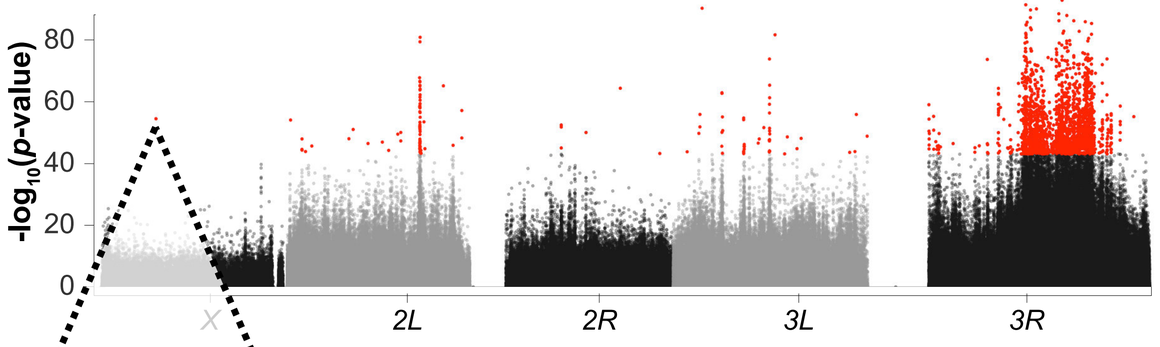
Chr	Pos	AT_Mau_14_01	DE_Mun_14_31	DK_Kar_14_41	ES_Gim_14_34	FI_Aka_14_36	PT_Rec_14_33	RU_Ale_15_55	TR_Yes_14_04	UA_Uma_15_9
2L	16816	0.081	0.0	0.091	0.022	0.0	0.021	0.135	0.0	0.021
2L	17549	0.0	0.028	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2L	17582	0.405	0.368	0.4	0.455	0.429	0.163	0.385	0.167	0.519
2L	35621	0.25	0.533	0.514	0.413	0.519	0.19	0.356	0.357	0.448
2L	39078	0.0	0.0	0.073	0.0	0.0	0.0	0.0	0.0	0.0
2L	59476	0.031	0.021	0.023	0.0	0.082	0.0	0.0	0.0	0.027
2L	66172	0.068	0.132	0.0	0.023	0.0	0.0	0.073	0.065	0.071
2L	83056	0.311	0.159	0.49	0.222	0.304	0.152	0.273	0.209	0.491
2L	92751	0.061	0.0	0.0	0.043	0.188	0.02	0.306	0.045	0.154
2L	100908	0.0	0.08	0.0	0.059	0.0	0.024	0.0	0.0	0.022
2L	112460	0.0	0.021	0.0	0.121	0.0	0.127	0.0	0.029	0.0
2L	117028	0.0	0.0	0.026	0.0	0.05	0.0	0.0	0.0	0.0
2L	129746	0.051	0.043	0.0	0.048	0.0	0.0	0.0	0.037	0.016
2L	130390	0.0	0.0	0.0	0.019	0.0	0.0	0.0	0.0	0.0
2L	150055	0.036	0.0	0.0	0.0	0.031	0.0	0.158	0.111	0.0

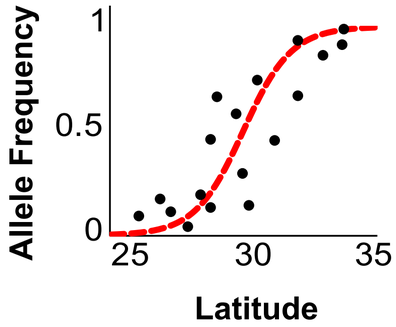
metadata

sampleId	collectionDate	lat	long	bio1	bio12
AT_Mau_14_01	2014/07/20	48,37	15,56	88	688
DE_Mun_14_31	2014/06/19	48,18	11,61	80	911
DK_Kar_14_41	2014/11/25	55,95	10,21	78	605
ES_Gim_14_34	2014/10/20	41,62	0,62	152	420
FI_Aka_14_36	2014/07/25	61,10	23,52	39	595
PT_Rec_14_33	2014/09/26	41,15	-8,41	145	1161
RU_Ale_15_55	2015/09/15	56,41	38,72	43	630
TR_Yes_14_04	2014/10/23	40,23	32,26	118	449
UA_Uma_15_9	2015/09/26	48,75	30,21	72	617

genome-wide patterns

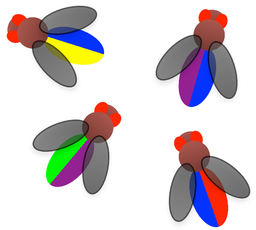


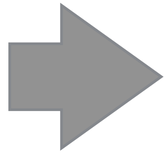




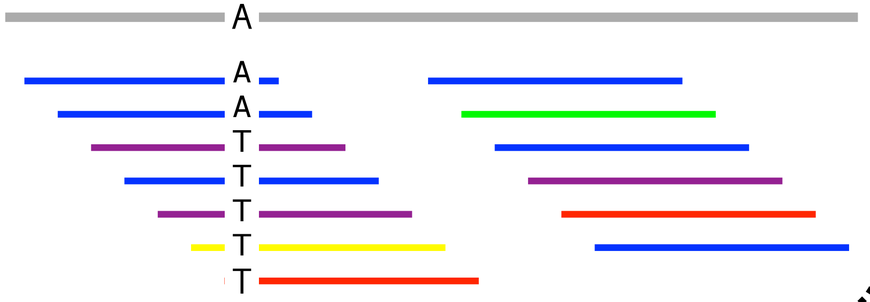
logistic regression:
allele-frequ. ~ latitude

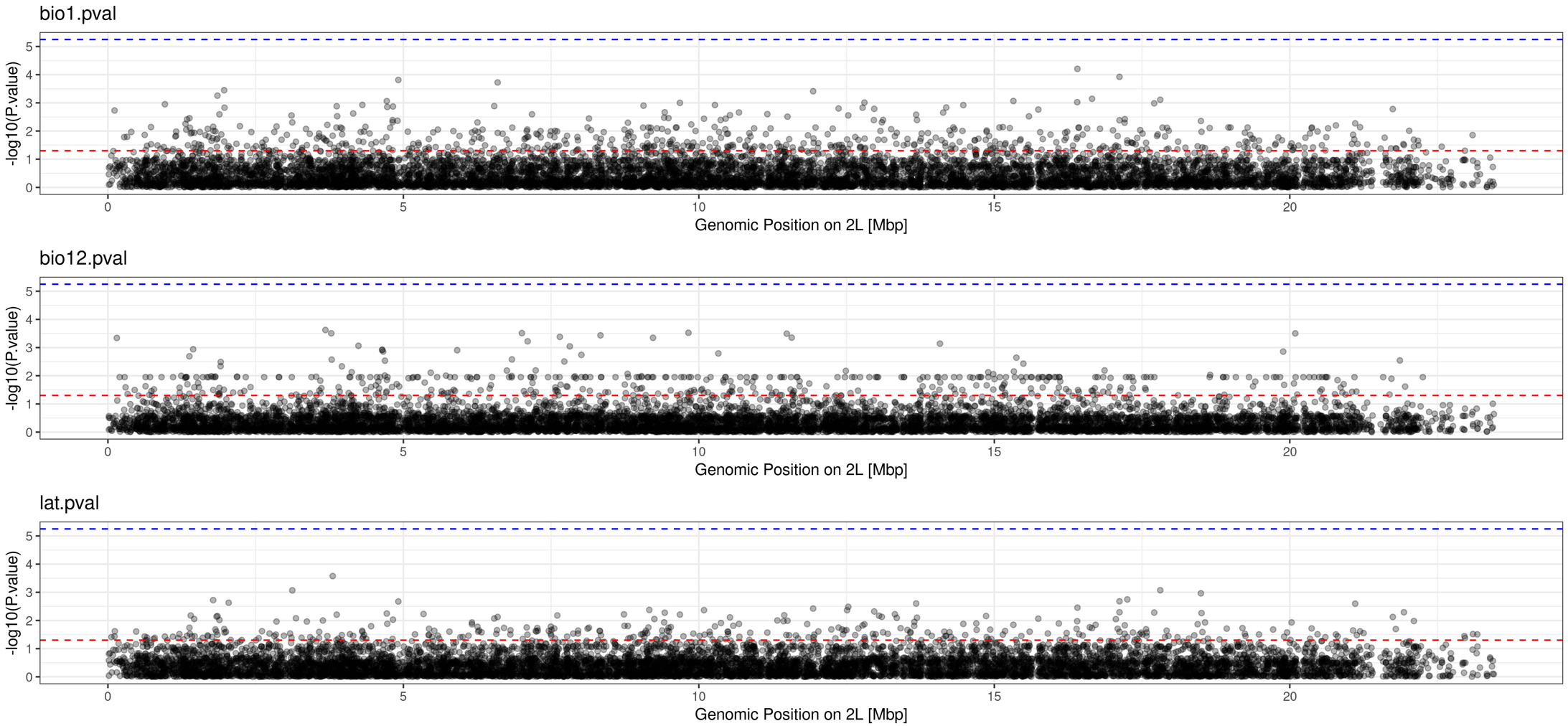
Pool-Seq





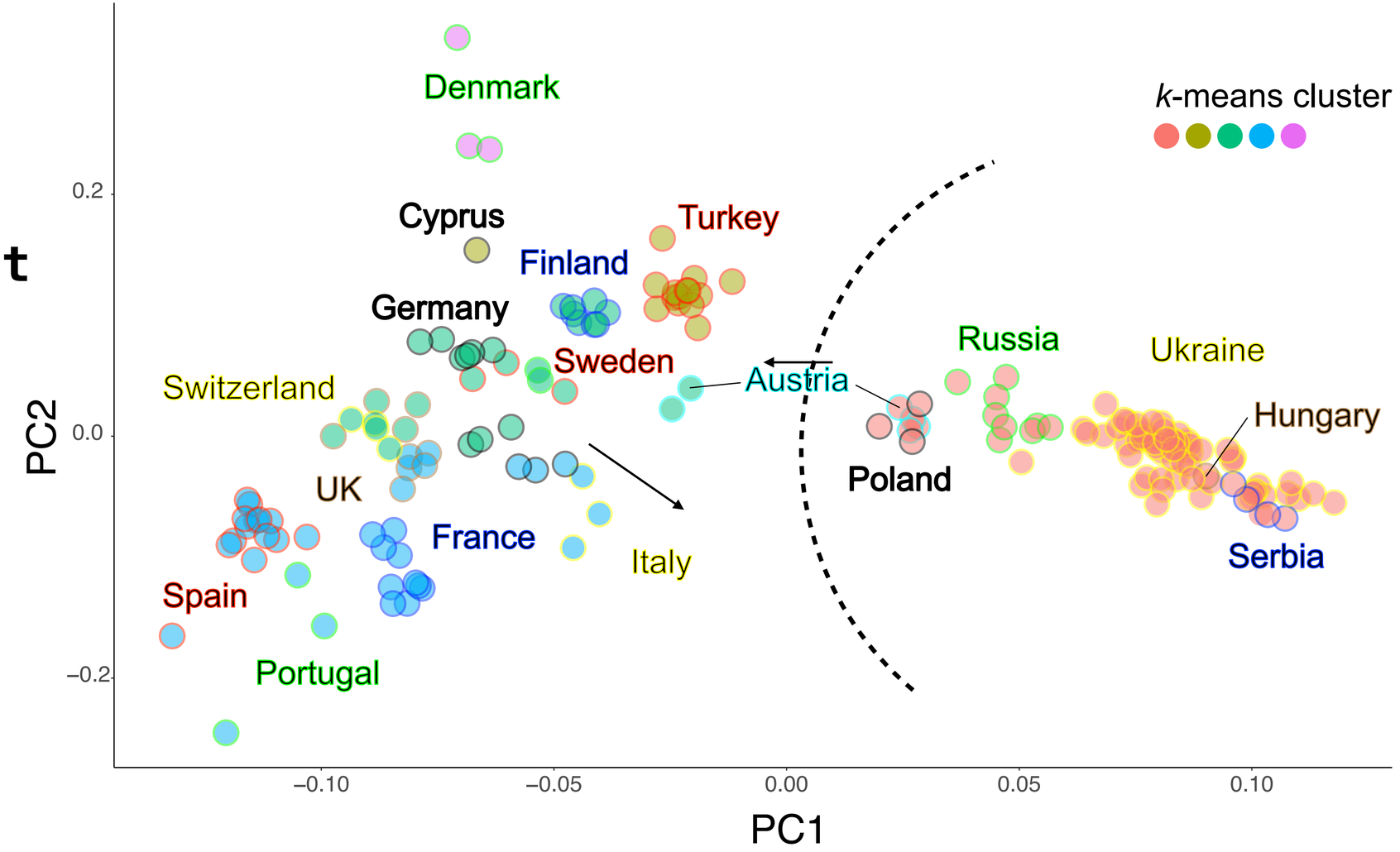
reference-genome




bio1.pval
-log10(P.value)
0
1
2
3
4
5
0
5
10
15
20
Genomic Position on 2L [Mbp]
bio12.pval
-log10(P.value)
0
1
2
3
4
5
0
5
10
15
20
Genomic Position on 2L [Mbp]
lat.pval
-log10(P.value)
0
1
2
3
4
5
0
5
10
15
20
Genomic Position on 2L [Mbp]

population structure?

- **20,000 SNPs**
- **Principal component analysis** (PCA) based on allele-frequencies
- 5 Cluster
- **east-west** distribution



Kapun *et al.* (2020) MBE
Kapun *et al.* (2021) MBE