

Current challenge in landscape genomics: What about the environmental counterpart of high-throughput genomic data?

stephane.joost@epfl.ch

Laboratory of Geographic Information Systems (LASIG, EPFL)
Geographic Information Research and Analysis for Public Health (GIRAPH)
Unit of Population Epidemiology, (UEP, HUG)

University & Lab

- EPFL, Lausanne, Switzerland
- School of Architecture, Civil and Environmental Engineering (ENAC)
- Institute of Environmental Engineering (IIE)
- Analysis of the relationship between living organisms and their environment
- Use of Geographic Information Systems and spatial statistics to analyse health data (spatial epidemiology) and genetic resources (landscape genomics)



Introduction

PERSPECTIVES

Science, 2010

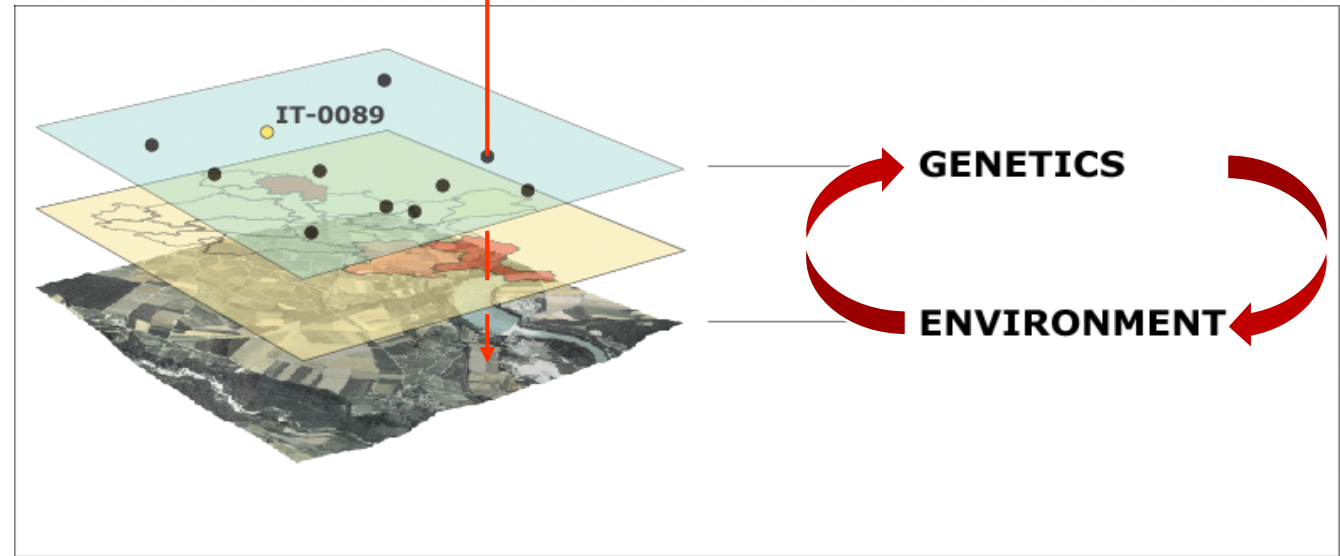
ECOLOGY

Time to Tap Africa's Livestock Genomes

Olivier Hanotte,¹ Tadelle Dessie,² Steve Kemp³

Fortunately, the fields of genetics and genomics (3–5) offer a new start for the sustainable improvement of African livestock productivity. **Landscape genomics links genome-wide information to geo-environmental resource analysis to identify potentially valuable genetic material.** Typically, researchers will perform a genome-wide scan on a number of animals from populations living in different habitats or across an ecological cline (from dry to wet areas, for instance).

Spatial coincidence



Landscape genomics


Link genome-wide information with geo-environmental data by means of correlative approaches

Introduction

Individuals			Genetic data													Environmental variables																
1	famid	animalid	DARJMP29_allele2_137	DARJMP29_allele2_138	DARJMP29_allele2_140	DARJMP29_allele2_141	DARJMP29_allele2_142	DARJMP29_allele2_143	DARJMP29_allele2_144	DARJMP29_allele2_145	DARJMP29_allele2_146	DARJMP29_allele2_147	DARJMP29_allele2_148	DARJMP29_allele2_149	DARJMP29_allele2_150	DARJMP29_allele2_151	DARJMP29_allele2_152	DARJMP29_allele2_153	DARJMP29_allele2_154	DARJMP29_allele2_155	DARJMP29_allele2_156	wndjan	altitude	wndfeb	wndmar	wndapr						
1044	PL-4005	QAPLPOM25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	22	4.6	5	4.4
1045	PL-4005	QAPLPOM26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	22	4.6	5	4.4
1046	PL-4006	QAPLPOM01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1047	PL-4006	QAPLPOM15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1048	PL-4006	QAPLPOM24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1049	PL-4007	QAPLPOM05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	250	4.8	5	4.5
1050	PL-4007	QAPLPOM16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	250	4.8	5	4.5
1051	PL-4008	QAPLPOM09	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1052	PL-4008	QAPLPOM19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1053	PL-4008	QAPLPOM20	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1054	PL-4009	QAPLPOM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.5	87	5	5.2	4.6
1055	PL-4009	QAPLPOM21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.5	87	5	5.2	4.6
1056	PL-4010	QAPLPOM08	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.4	208	4.9	5.1	4.5

- Mitton (1977) first had the idea to correlate the frequency of alleles with an environmental variable to look for signatures of selection in Ponderosa pine
- Multiple parallel logistic regressions (Joost et al. 2007), MatSAM, now Sambada (Stucki et al. 2016)

Theor. Appl. Genet. 51, 5-13 (1977)



© by Springer-Verlag 1977

Observations on the Genetic Structure and Mating System of Ponderosa Pine in the Colorado Front Range

J.B. Mitton, Y.B. Linhart, J.L. Hamrick and J.S. Beckman
Department of Environmental, Population, and Organismic Biology, University of Colorado, Boulder, Colorado (U.S.A.)

Summary. Variation of peroxidase enzymes is analyzed both in mature needle tissue and in open-pollinated seedling families of ponderosa pine, *Pinus ponderosa*, and is identified as being controlled by a single Mendelian locus. Variation at this locus, analyzed in 1,386 individuals, is used in the analysis of population differentiation and the mating system. Significant variation of gene frequencies is detected over distances of several hundred meters, and is found to be associated with slopes of different aspects. Ponderosa pine is wind-pollinated, and an analysis of the mating system indicates that the level of outcrossing is greater than 90%. Selection specific for different environments is evidently strong enough to overcome the homogenizing force of migration and produce population fissuring in ponderosa pine.

Key Words: Selection - Migration - Peroxidase - Ponderosa Pine

Molecular Ecology (2007) 16, 3955-3969

doi: 10.1111/j.1365-294X.2007.03442.x

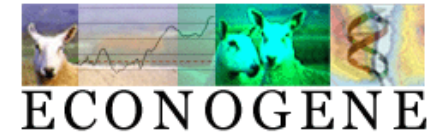
A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation

S. JOOST,†A. BONIN,‡M. W. BRUFORD,§L. DESPRÉS,‡C. CONORD,‡G. ERHARDT¶ and P. TABERLET†**

*Istituto di Zootechnica, Università Cattolica del S. Cuore, via E. Parmense 84, 29100 Piacenza, Italy, †Laboratoire de Systèmes d'Information Géographique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland, ‡Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 09, France; §Cardiff School of Biosciences, Main Building, Museum Avenue, Cardiff CF10 3TL, UK, ¶Department of Animal Breeding and Genetics, Justus-Liebig-University of Giessen, Ludwigstrasse 21B, 35390 Giessen, Germany

Introduction

- When I started computing association models in landscape genomics...
- 2005: Common frog – 302 markers (AFLPs) x 1 env. var (altitude) = 302 models
- 2007: Sheep & goats – 750 markers (microsats, SNPs, AFLPs) x 120 env. var. (CRU)= 90'000 models
- ...
- 2016: Sheep & goats – Whole Genome Sequence Data: 35 mio SNPs x 100 env. var. (over 3 billion models)




- Gradual increase of the resolution of genomic information (DNA resolution in base pairs)
- Advent of High-throughput genomic data, new avenues for research

Introduction

Individuals			Genetic data														Environmental variables								
1	famid	animalid	DARJMP29_allele2_137	DARJMP29_allele2_138	DARJMP29_allele2_141	DARJMP29_allele2_142	DARJMP29_allele2_143	DARJMP29_allele2_144	DARJMP29_allele2_145	DARJMP29_allele2_146	DARJMP29_allele2_147	DARJMP29_allele2_148	DARJMP29_allele2_149	DARJMP29_allele2_150	DARJMP29_allele2_151	DARJMP29_allele2_152	DARJMP29_allele2_153	DARJMP29_allele2_154	DARJMP29_allele2_155	DARJMP29_allele2_156	wndjan	altitude	wndfeb	wndmar	wndapr
1044	PL-4005	QAPLPOM25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	22	4.6	5	4.4
1045	PL-4005	QAPLPOM26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	22	4.6	5	4.4
1046	PL-4006	QAPLPOM01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1047	PL-4006	QAPLPOM15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1048	PL-4006	QAPLPOM24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	153	4.8	4.9	4.3
1049	PL-4007	QAPLPOM05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	250	4.8	5	4.5
1050	PL-4007	QAPLPOM16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.3	250	4.8	5	4.5
1051	PL-4008	QAPLPOM09	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1052	PL-4008	QAPLPOM19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1053	PL-4008	QAPLPOM20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	5.2	166	4.8	5	4.4
1054	PL-4009	QAPLPOM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.5	87	5	5.2	4.6
1055	PL-4009	QAPLPOM21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.5	87	5	5.2	4.6
1056	PL-4010	QAPLPOM08	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5.4	208	4.9	5.1	4.5

Theor. Appl. Genet. 51, 5-13 (1977)



© by Springer-Verlag 1977

Observations on the Genetic Structure and Mating System of Ponderosa Pine in the Colorado Front Range

J.B. Mitton, Y.B. Linhart, J.L. Hamrick and J.S. Beckman
Department of Environmental, Population, and Organismic Biology, University of Colorado, Boulder, Colorado (U.S.A.)

Summary. Variation of peroxidase enzymes is analyzed both in mature needle tissue and in open-pollinated seedling families of ponderosa pine, *Pinus ponderosa*, and is identified as being controlled by a single Mendelian locus. Variation at this locus, analyzed in 1,386 individuals, is used in the analysis of population differentiation and the mating system. Significant variation of gene frequencies is detected over distances of several hundred meters, and is found to be associated with slopes of different aspects. Ponderosa pine is wind-pollinated, and an analysis of the mating system indicates that the level of outcrossing is greater than 90%. Selection specific for different environments is evidently strong enough to overcome the homogenizing force of migration and produce population fissuring in ponderosa pine.

Key Words: Selection - Migration - Peroxidase - Ponderosa Pine

Molecular Ecology (2007) 16, 3955-3969 doi: 10.1111/j.1365-294X.2007.03442.x

A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation

S. JOOST,†A. BONIN,‡M. W. BRUFORD,§L. DESPRÉS,‡C. CONORD,‡G. ERHARDT¶ and P. TABERLET†**

*Istituto di Zootecnica, Università Cattolica del S.Cuore, via E. Parmense 84, 29100 Piacenza, Italy, †Laboratoire de Systèmes d'Information Géographique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland, ‡Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 09, France; §Cardiff School of Biosciences, Main Building, Museum Avenue, Cardiff CF10 3TL, UK, ¶Department of Animal Breeding and Genetics, Justus-Liebig-University of Giessen, Ludwigstrasse 21B, 35390 Giessen, Germany

The environmental counterpart of high-genomic resolution

- With environmental variables, one can increase the number of variables of different sources
- What not necessarily provides additional information
- Because of common information often shared by different climate variables for instance (redundancy between altitude, temperature, precipitation)

- The main interest is in increasing spatial resolution of the data
- **To extract at best the information likely to be produced by the use of high-throughput genomic data in landscape genomics**

Unbalanced situation

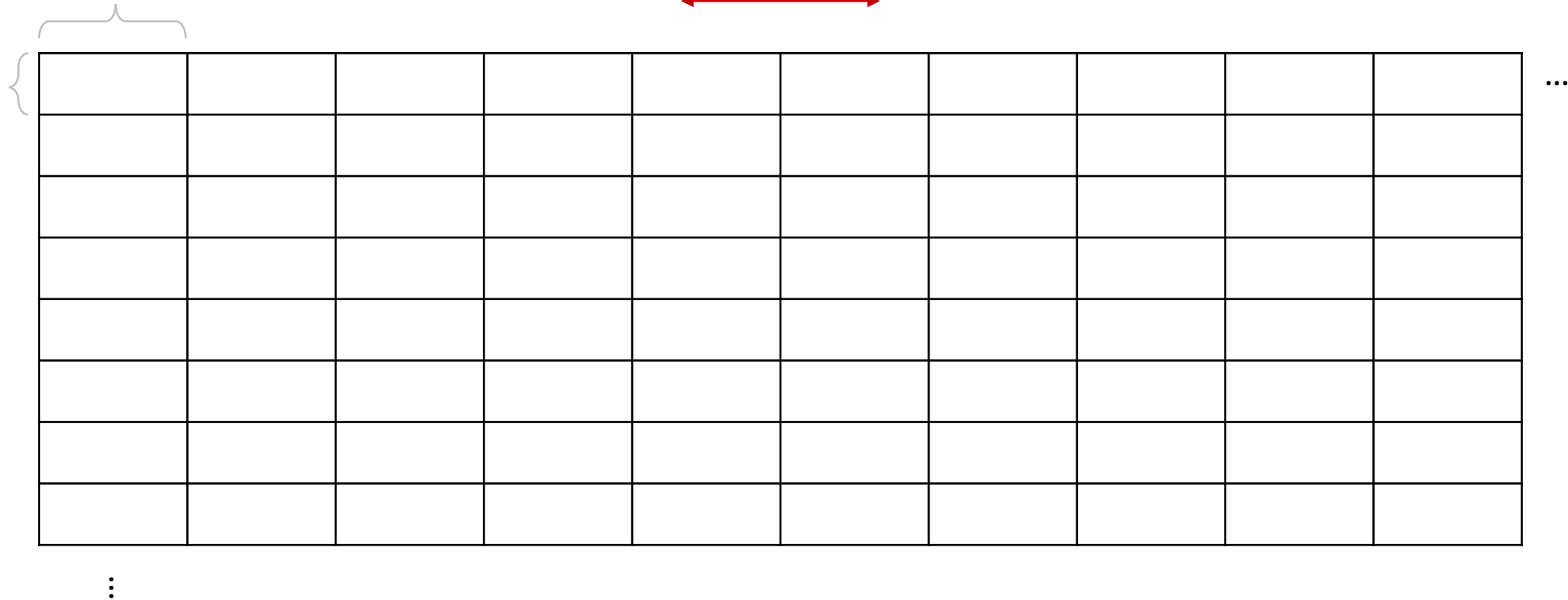
Geo-environmental data

Spatial resolution



Genomic data

Genomic resolution



Improving the resolution of geo-environmental data

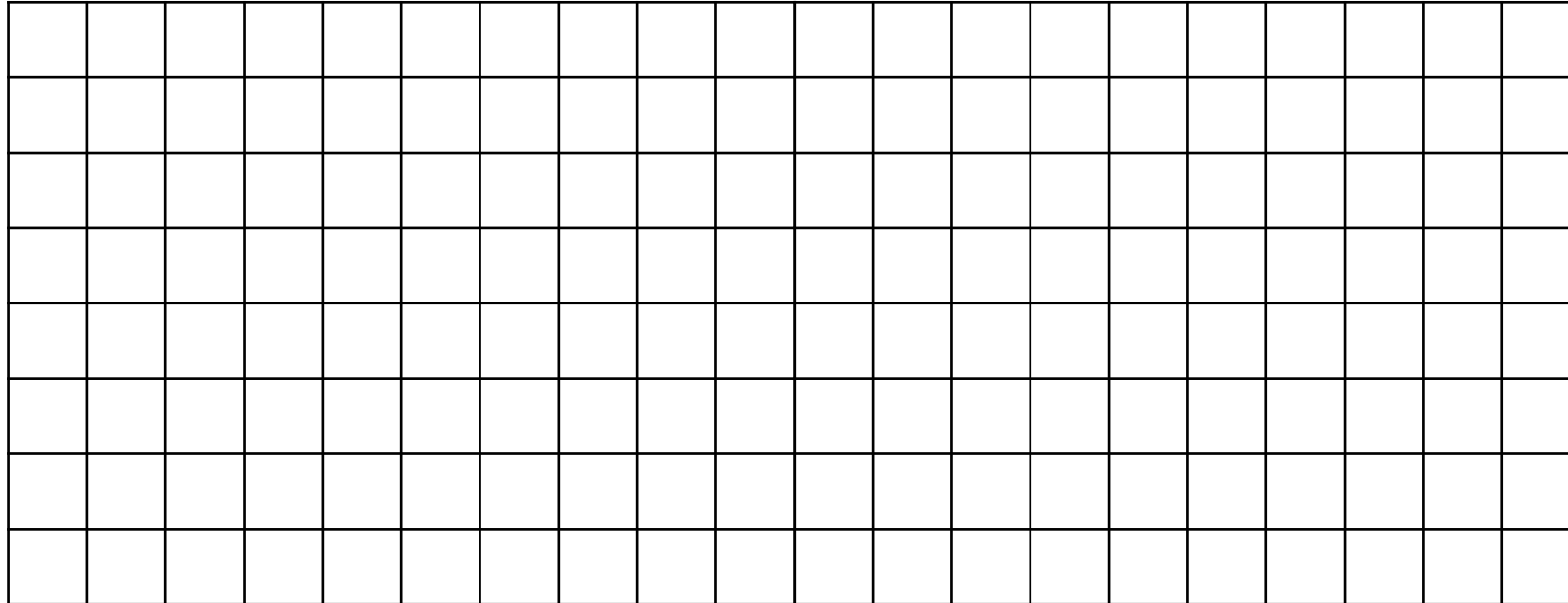
Geo-environmental data

Spatial resolution



Genomic data

Genomic resolution



Increasing the spatial resolution of environmental data

- There are two sub-topics:

1. Increasing the spatial resolution of existing data. There are plenty of geo-environmental data publicly available but often their spatial resolution is coarse and these data better fit large scale studies with sparse distribution of sampled individuals.

Downscaling (Enke & Spekat, Climate Research, 1997)

2. Producing new environmental variables with high or very high resolution, often at locations not covered by existing geo-environmental variables, or where spatial resolution is too coarse to fit high density sampling in a small area (local scale)
 - a) Creation of High resolution environmental variables from existing Digital Elevation Models (DEMs)
 - b) Processing of Very High Resolution (VHR) environmental variables from DEMs acquired by means of helicopters equipped with a LIDAR system or by UAVs (Unmanned Automated Vehicles or drones)

a) Existing DEMs to produce high resolution variables

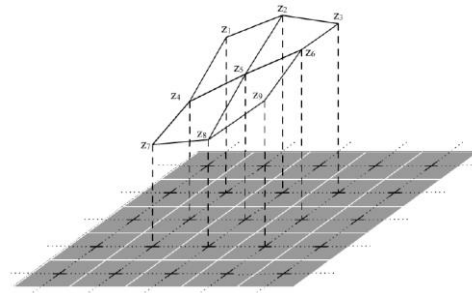
- Nextgen project (EU FP7 2010-2014) investigated local adaptation of sheep and goats in Morocco
- WGS data for 320 individuals carefully sampled across several contrasted environmental conditions
- Best environmental data available: Worldclim/Bioclim with 1km² spatial resolution: not sufficient
- We used a DEM produced on the basis of Shuttle Radar Topography Mission (SRTM) data (radar interferometry) with **90m²** spatial resolution (better quality than Aster - 30m²)
- To produce several DEM-derived environmental variables

DEM-derived variables

Zevenbergen & Thorne (1987) Quantitative analysis of land surface topography

Primary attributes

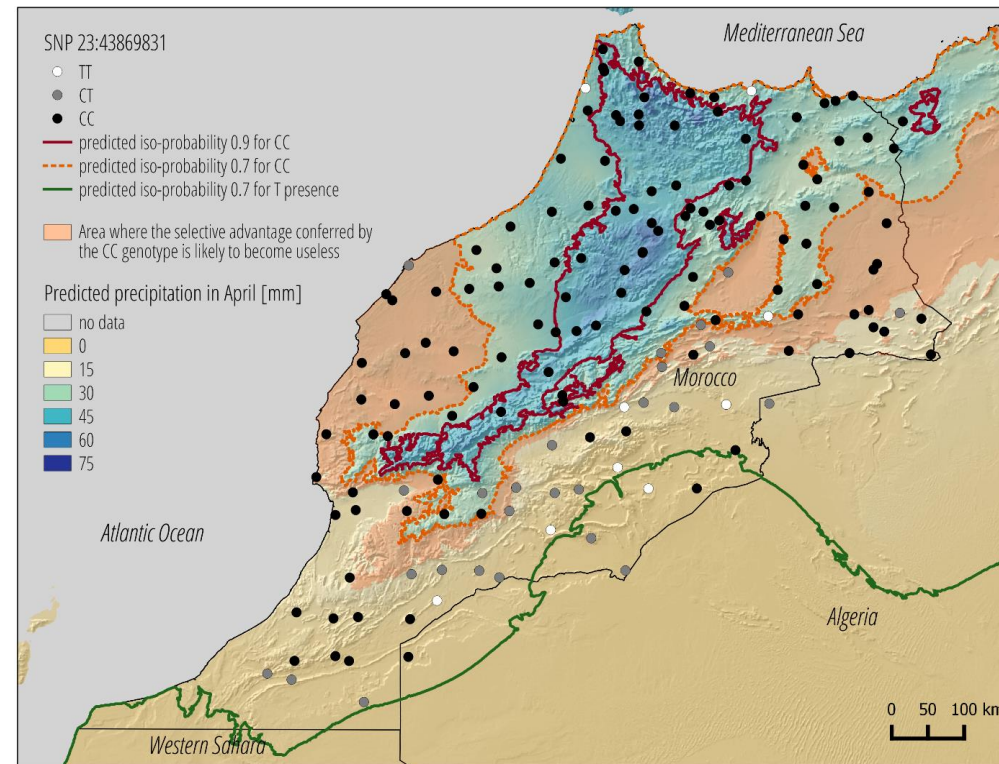
- Aspect
- Slope
- Curvature



Second derivatives

- Morphometric Protection Index
- Sky View Factor
- Vector Ruggedness Measure
- Total Insolation
- Direct insolation
- Terrain Wetness Index
- Temperature
- Etc.

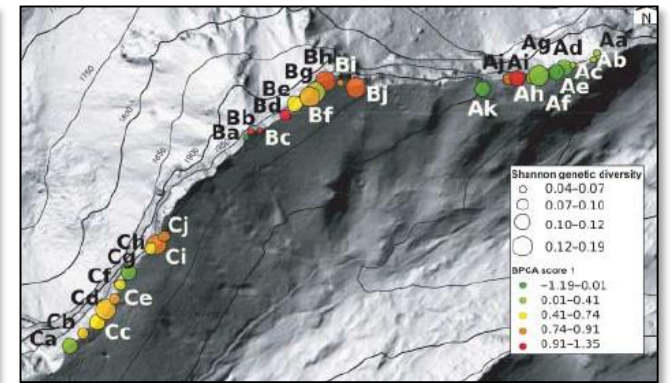
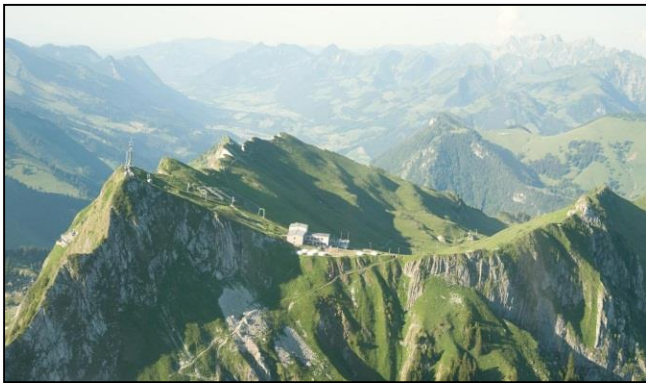
- Mainly related to solar radiation, light, humidity, temperature
- Main progress: better spatial resolution makes it possible to investigate more ecological/biological processes or phenomena (richer set of environmental descriptors)



Sampling locations in Morocco and Spatial Areas of Genotype Probability (SPAGs) based on SRTM-derived variables (Vajana et al. 2016)

b) Generate new DEMs to produce very high resolution variables

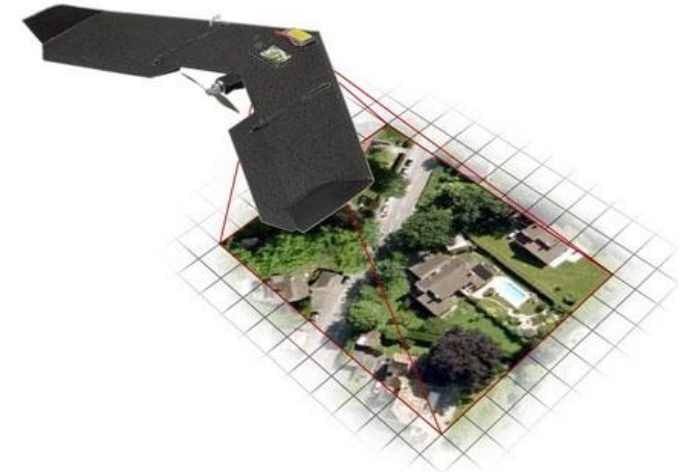
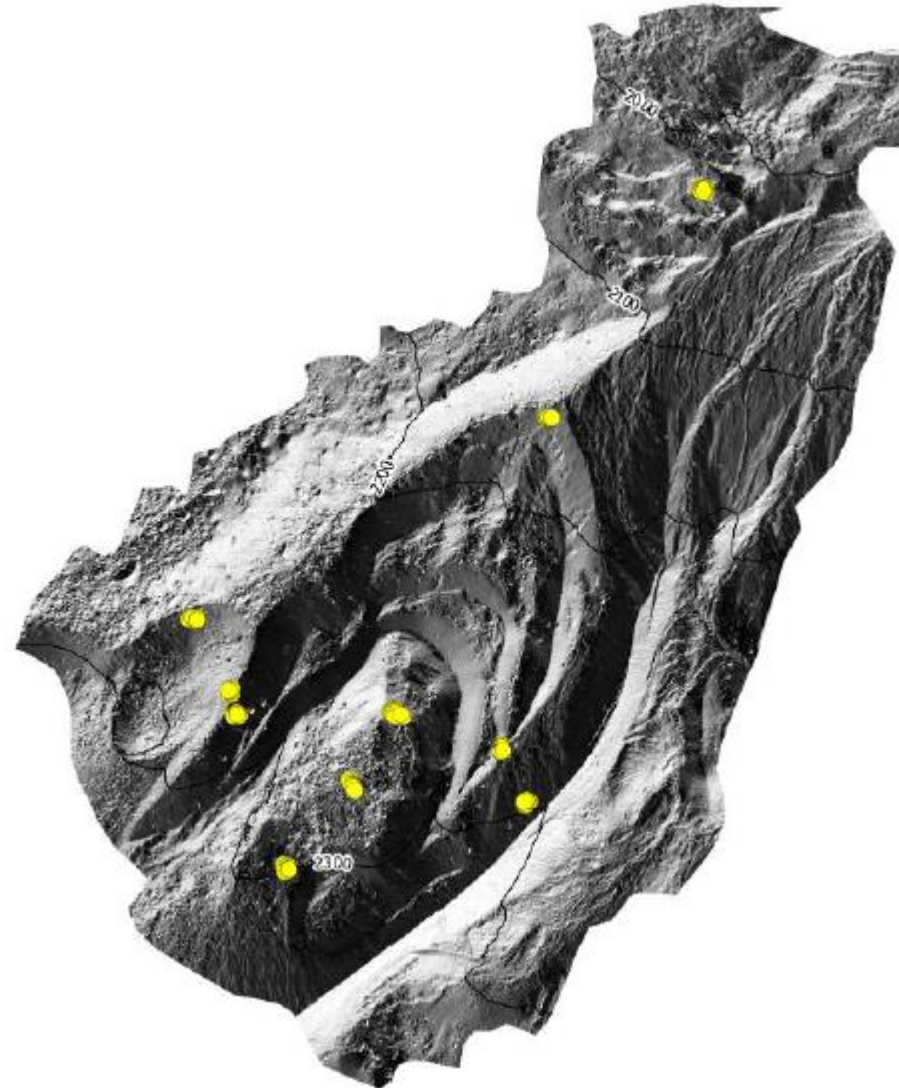
- The same types of variables can be produced starting from scratch and providing much finer spatial resolutions
 - When existing DEMs show a too broad resolution compared with an existing sampling density
 - And when the biological models studied require a more accurate description of their local environmental conditions (typically plants)



Two possible options for data acquisition

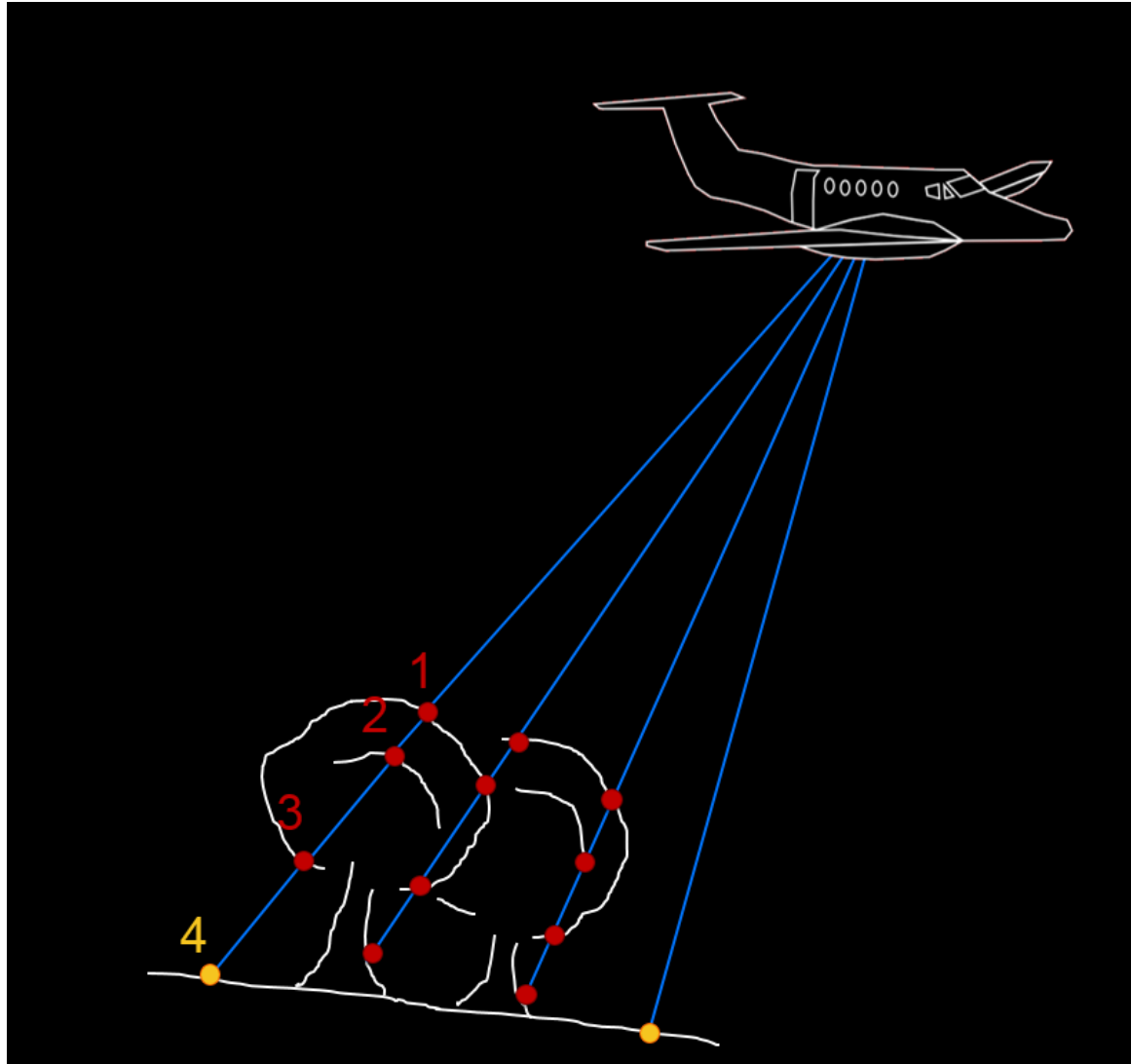


Helicopter - LIDAR



UAV or drone – IMAGE MATCHING

LIDAR (Light Detection and Ranging)



- pulses of light energy using a laser sent to the ground
- measure of how long it takes for the pulse to return
- 8-12 points (=altitude) per square meter

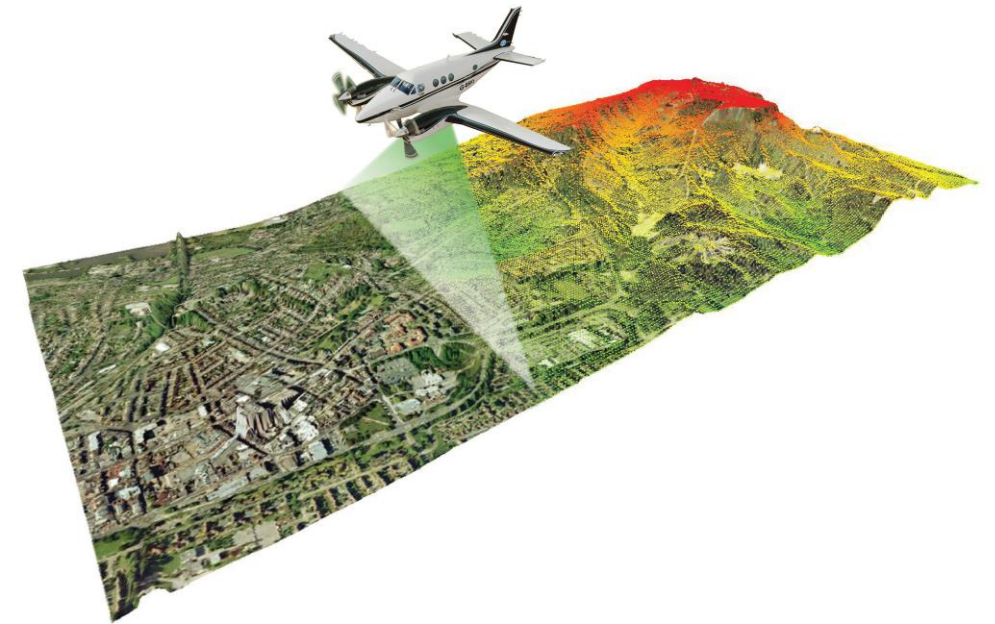
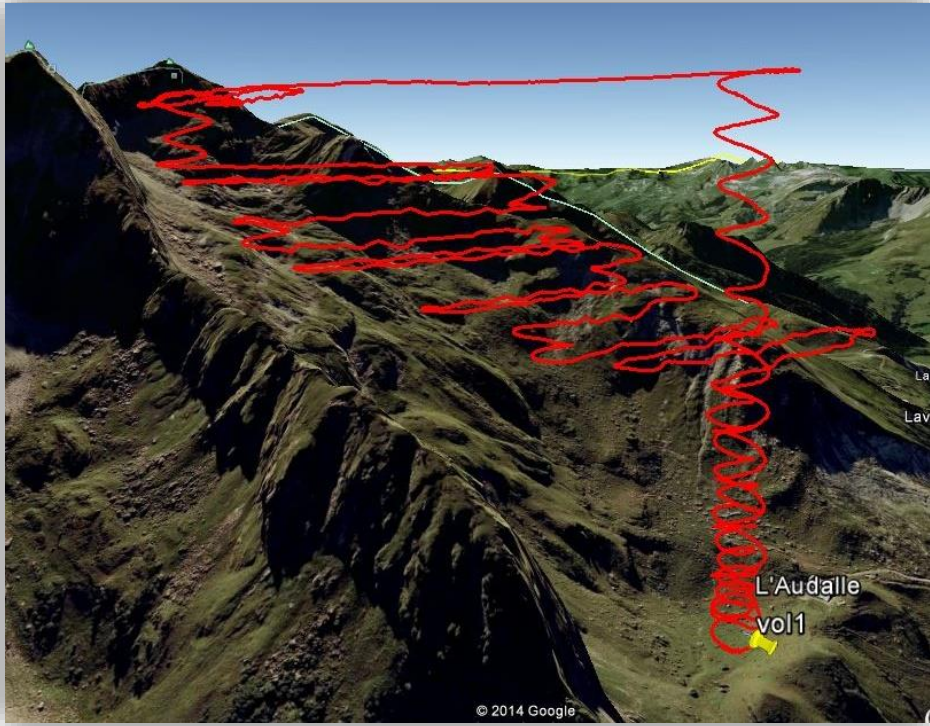
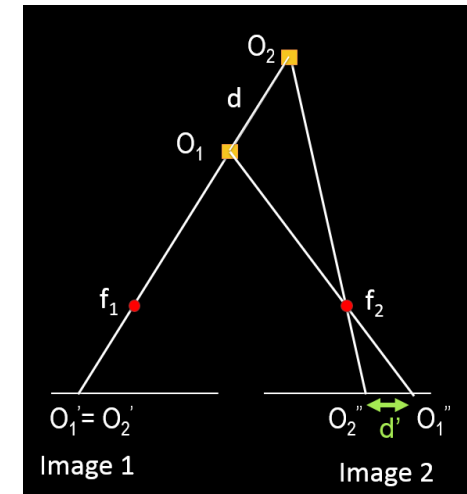
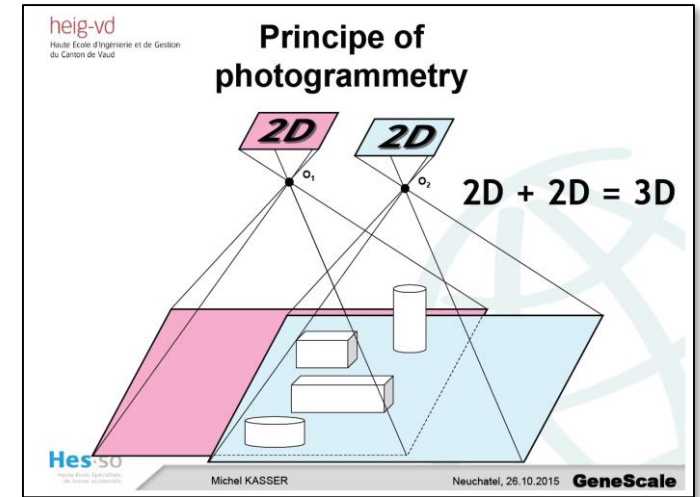
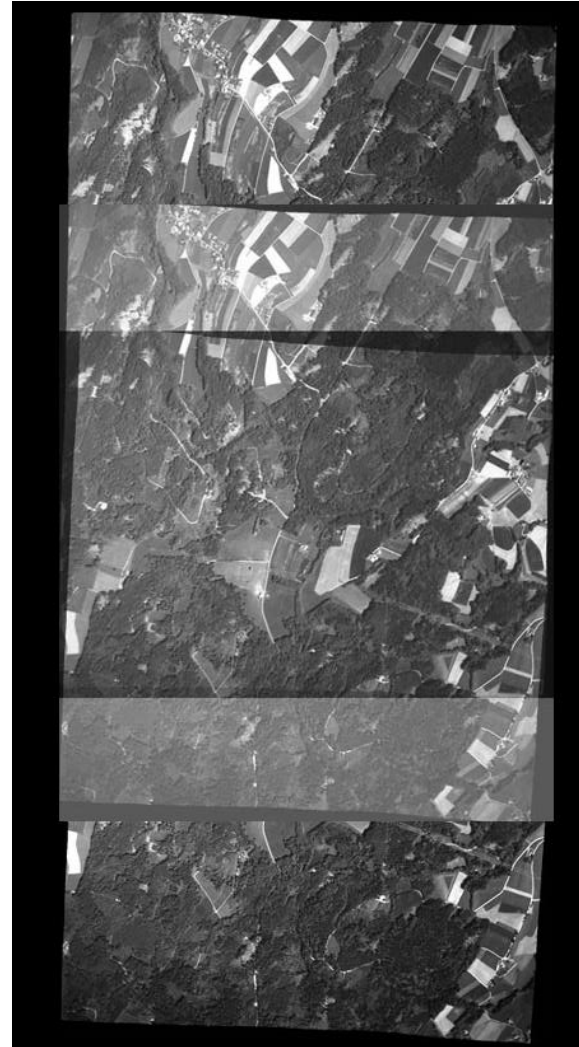


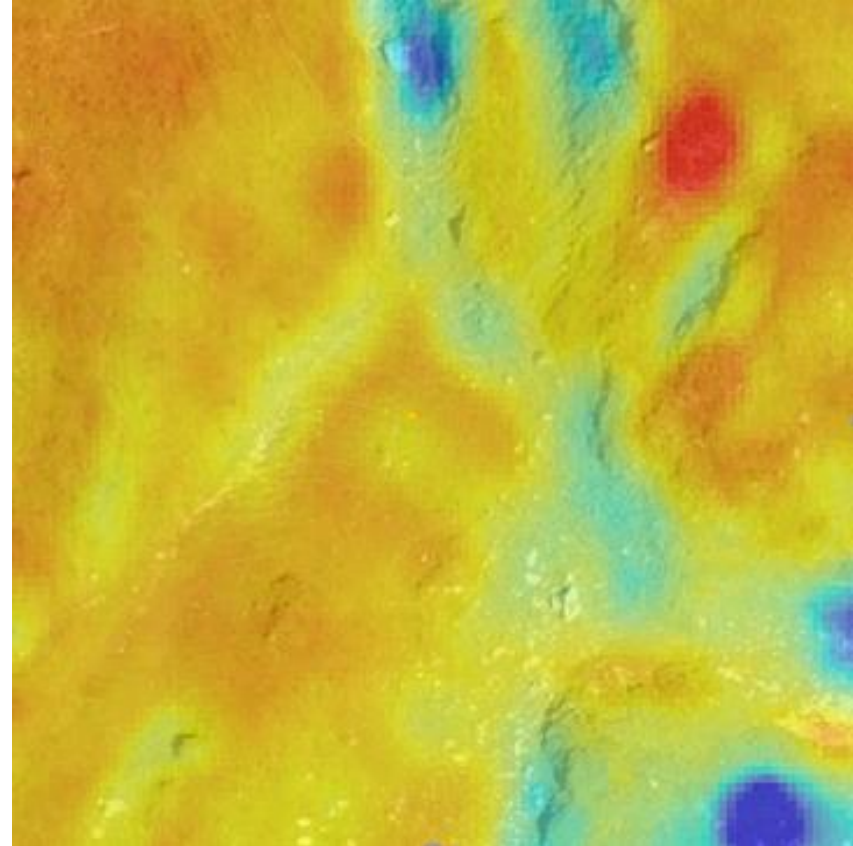
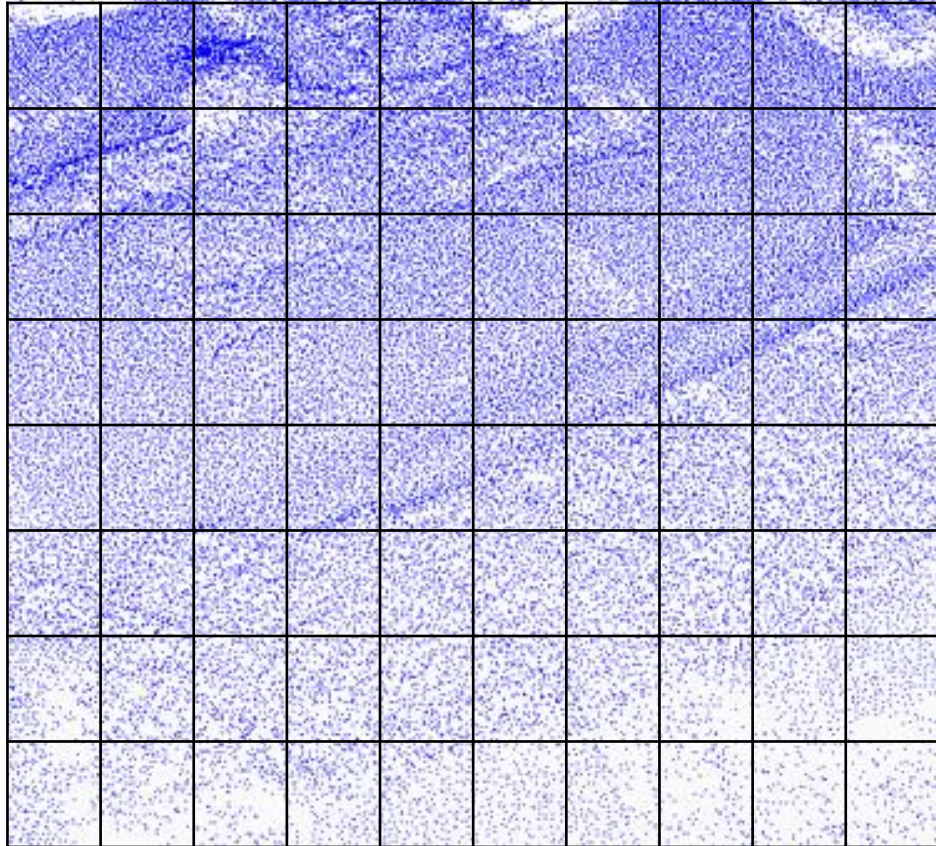
Image matching (stereophotogrammetry)



- Many overlapping images
- 60-100 points (=altitude) per square meter



Point cloud to interpolated regular grid



Spatial resolution of VHR DEMs and derived variables

Model	Helicopter/plane
Spatial resolution	20cm
Vert. accuracy	<10cm
LIDAR	

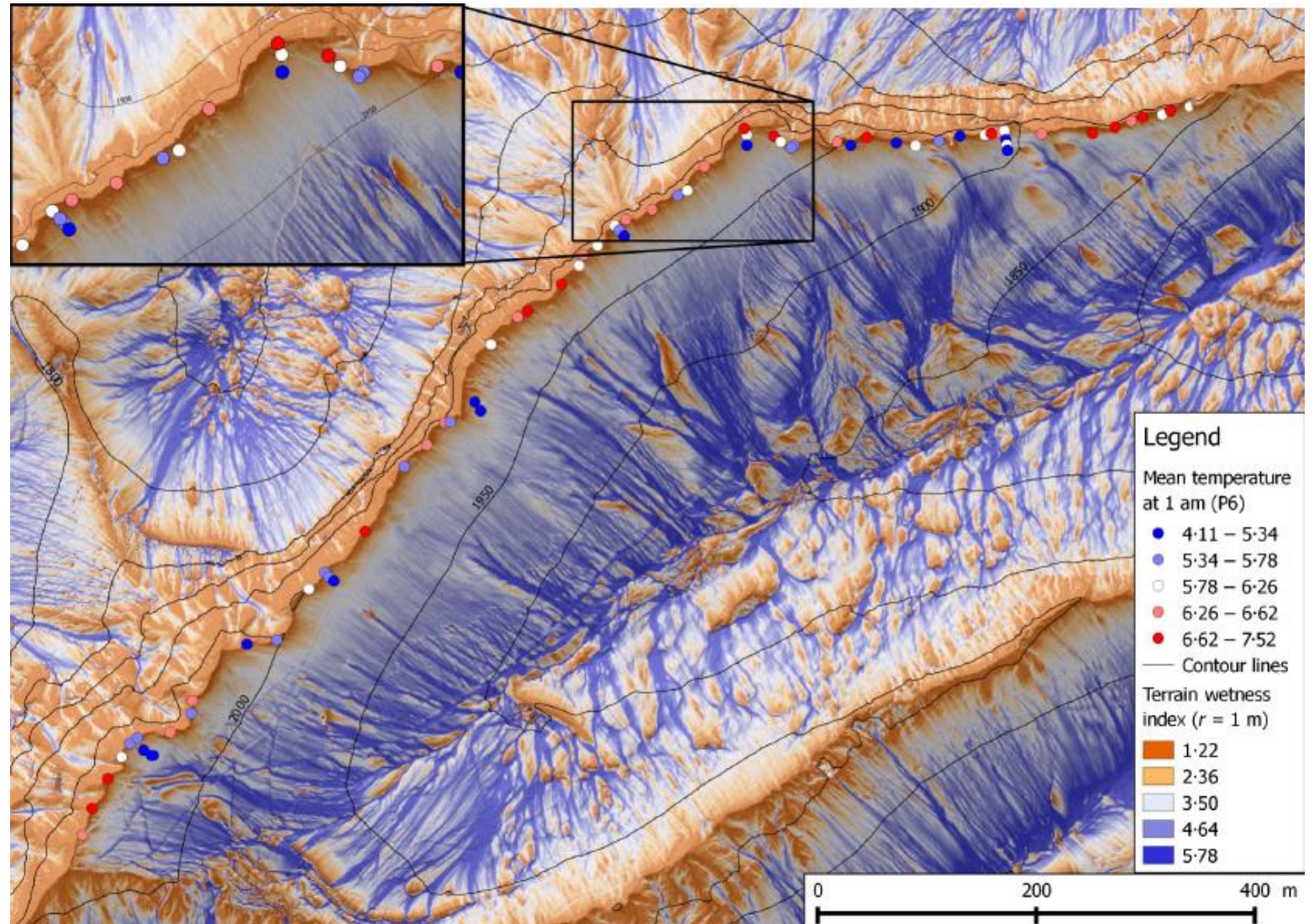
- Large areas covered – ok for solar and hydrology-related variables (shade, total radiation, soil temperature estimation, wetness, etc.)

Model	UAV
Spatial resolution	4cm
Vert. accuracy	≈50cm
IMAGE MATCHING	

Much smaller areas covered (limit = UAV's autonomy, ~30 min) – does not enable calculation of solar or hydrology-related variables: often we do not have the surrounding relief (too far away)

Ecological relevance of DEM's derived variables

- Important question: are these derived variables ecologically relevant?
- Produce nice maps, but meaningful?
- Case study in the Swiss Prealps (Naye) to compare these variables with data recorded by sensors (temperature, humidity loggers) in the field
- Calculation of regression models between DEM-derived variables and measured variables at different seasons



Ecological relevance of DEM's derived variables

- Specific VHR DEM-derived variables show significant associations with climatic factors
- Spatial resolution of DEM-derived variables has a significant influence on models' strength, with coefficients of determination decreasing with coarser resolutions or showing an optimum for a specific resolution
- The results obtained support the relevance of using **multi-scale** DEM variables
- Provide surrogates for important variables like humidity, moisture, temperature: suitable alternative to direct measurements

Methods in Ecology and Evolution



Methods in Ecology and Evolution 2015, 6, 1373–1383

doi: 10.1111/2041-210X.12427

Very high-resolution digital elevation models: are multi-scale derived variables ecologically relevant?

Kevin Leempoel^{1*}, Christian Parisod², Céline Geiser², Lucas Daprà², Pascal Vittoz³ and Stéphane Joost¹

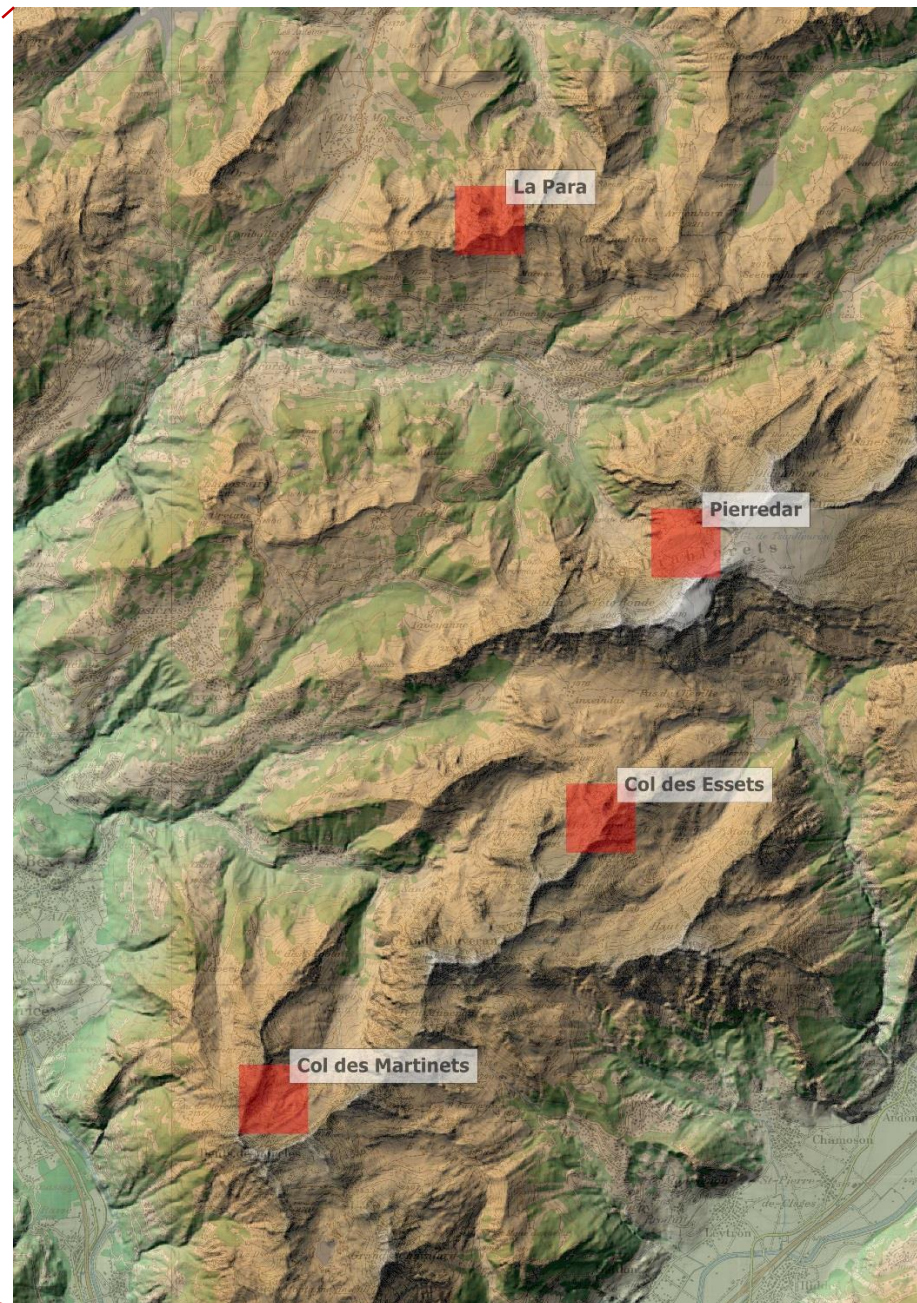
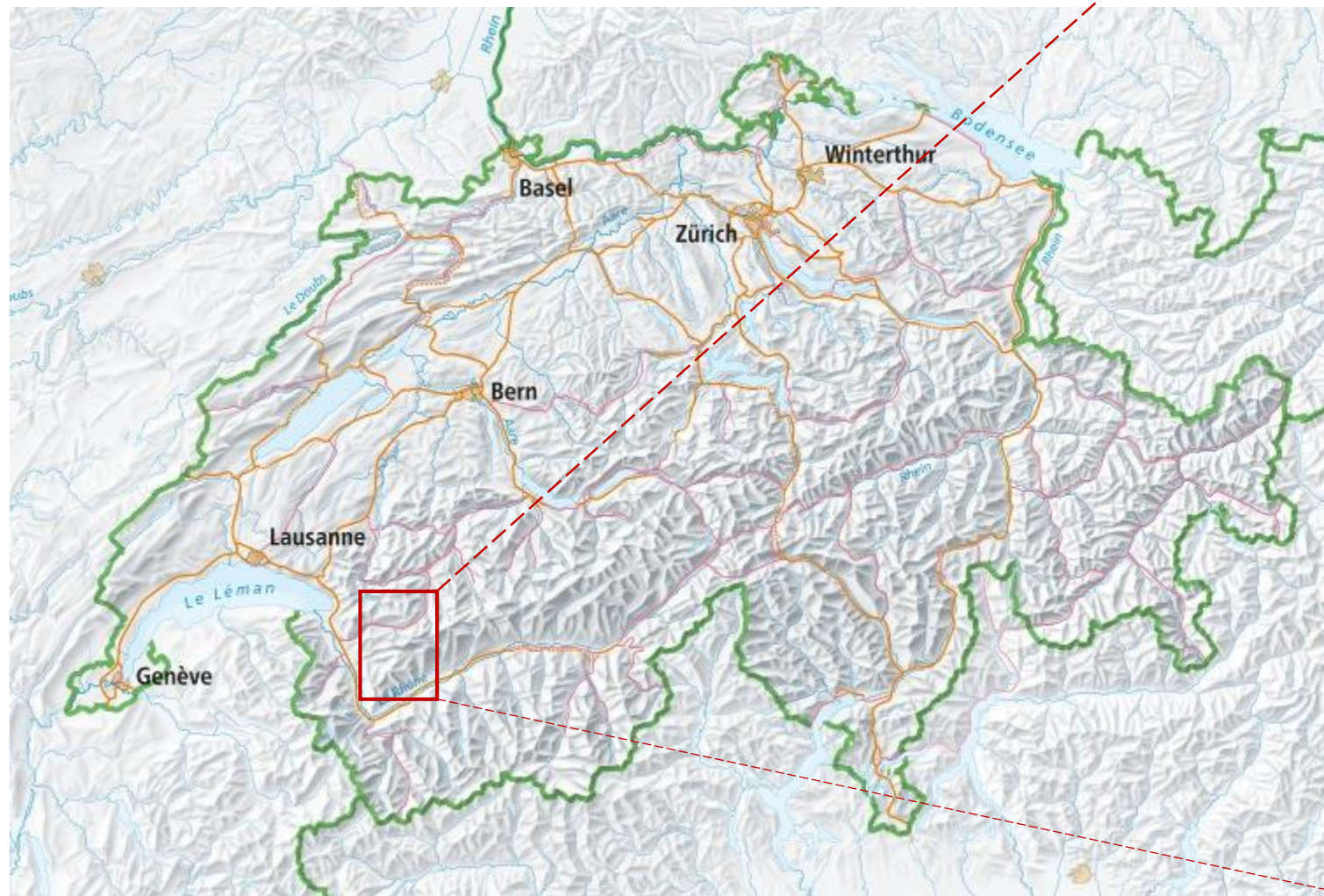
¹Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland; ²Laboratory of evolutionary botany, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland; and ³Institut des dynamiques de la surface terrestre, University of Lausanne, Géopolis, CH-1015 Lausanne, Switzerland

GENESCALE project (WSL, EPFL, UNINE, HEIG-VD)

- So let's implement a multi-scale landscape genomics study...
- And benefit from the simultaneous use of high-throughput genomic data and VHR environmental variables
- “Very high-resolution digital elevation models for **multi-scale** analysis in landscape genomics”
- Adaptation of *Arabidopsis thaliana* to its local environment in 4 study areas
- Opportunity to answer the question: “**at which spatial scale does natural selection operate?**”



Study areas



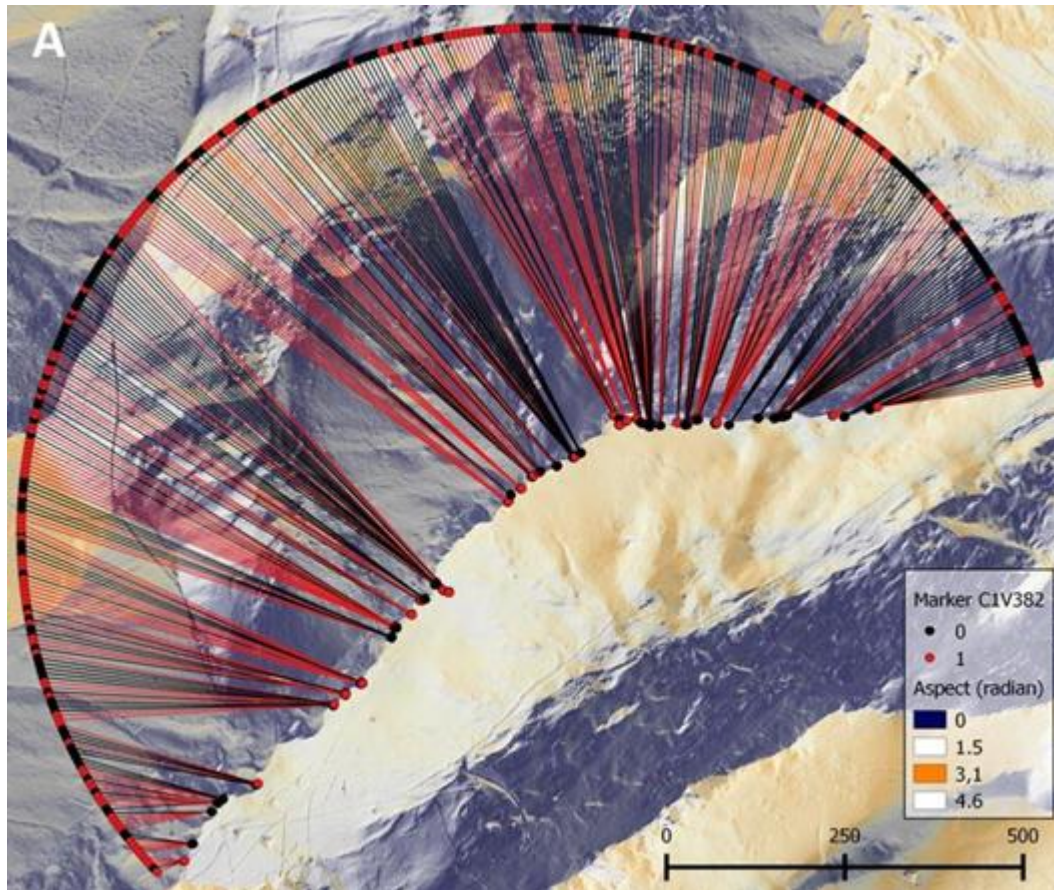
~400'000 SNPs x 4cm spatial resolution...

	No. of individuals	No. of SNPs	% of SNPs heterozygotes	Mean F_{is}
ALL	304	439'670	9.51*	0.29
Essets	70	329'946	8.96	0.33
Martinets	96	177'909	5.44	0.59
Pierredar	69	239'813	7.6	0.43
Para	69	434'110	17.65	-0.30

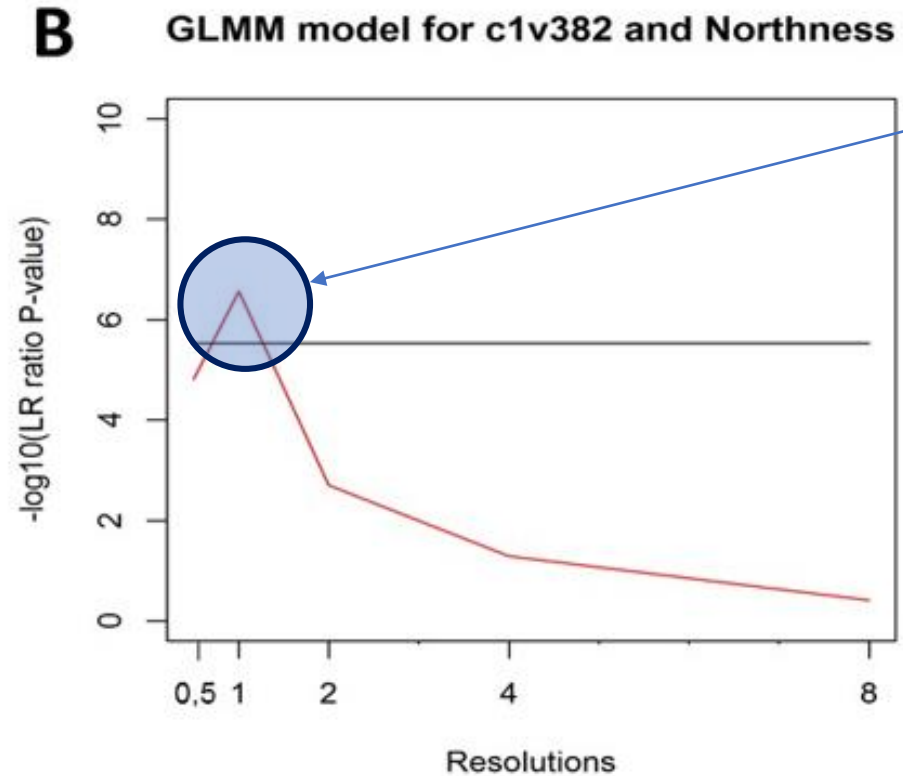


- More information on Friday, Symposium 16 «Genomics of adaptation», Room B, 12h30 : **Aude Rogivue** et al. Environmental factors driving local adaptation in the Alpine Brassicaceae *Arabis alpina*

Just a foretaste...



Spatial distribution of plant individuals along a ridge, red point showing locations where the marker of interest is present



Optimum with 1m resolution, for which Northness is most significantly associated with the genetic marker

Variation of the significance of association models between the genetic marker and Northness for different spatial resolutions

Conclusion

- This topic fully lies within the scope of scale issues discussed by John Wiens in 1989 and Simon Levin in 1992
- Wiens defined the notions of **extent** and **grain** of a study area
- They explained that the ability to detect patterns was a function of both the extent and the grain
- ... that the examination of ecological/biological phenomena require the study of how patterns change with the scale of description
- They mentioned the necessity to **quantify patterns** of variability in space and time, to understand how patterns change with scale
- ... the necessity to understand how information is transferred across scales
- Anticipated the role of “remote sensing, spatial statistics, and other methods...” to carry out these tasks
- Interesting to note that 25 years ago, Wiens and Levin described a theoretical framework we just started experimenting



J. A. Wiens (1989) Functional Ecology Vol. 3, No. 4, pp. 385-397



S.A. Levin (1992) Ecology, Vol. 73, No. 6, pp. 1493-1967

Conclusion

- What are the advantages of using very high resolution environmental variables in landscape genomics?
 - Make it possible to address phenomena at a local scale (e.g. range=1-2kms, grain=20m), or even enable landscape genomic studies for specific small species (micro-topography)
 - Enable **multi-scale analysis**, i.e. ...
 - Give the opportunity to address several possible ecological/biological processes «simultaneously»
 - **Empower high-throughput genomic data in spatial approaches: we know the details of DNA diversity, we need to compare it with the details of landscape diversity**
- What are the drawbacks?
 - Cost? e.g. UAV + navigation system and processing software = ~€ 17k, LIDAR more expensive
 - Still cheaper than high-throughput genomic data in a standard landscape genomic study (e.g. 5-10km² and 100 individuals)

Thank you for your attention!

Acknowledgments

Kevin Leempoel (EPFL, WSL, Stanford); Aude Rogivue (WSL), Michel Kasser, Stéphane Cretegy (HEIG-VD)

Felix Gugerli, Rimjhim Choudhury, Christian Parisod, François Felber

stephane.joost@epfl.ch