Genomic resources and their influence on the detection of the signal of positive selection in genome scans

S. Manel, Charles Perrier, Marine Pratlong, Laurent Abi-Rached, Julien Paganini, Pierre Pontarotti, D. Aurelle

MOLECULAR ECOLOGY

Genomic resources and their influence on the detection of the signal of positive selection in genome scans

Journal:	*Molecular Ecology*
Manuscript ID	MEC-15-1153.R1
Manuscript Type:	Invited Reviews and Syntheses
Date Submitted by the Author:	n/a
Complete List of Authors:	Manel, Stéphanie; CNRS, Centre d'Ecologie Fonctionnelle et Evolutive Perrier, Charles Pratlong, Marine Abi-Rached, Laurent Pontarotti, Pierre Paganini, Julien Aurelle, Didier
Keywords:	Adaptation, Ecological Genetics, Genomics/Proteomics, Landscape Genetics

Genomic resources and their influence on the detection of the signal of positive selection in genome scans

Manel S *, Perrier C*, Pratlong M †‡, Abi-Rached L**, Paganini J §, Pontarotti P ‡, Aurelle D †

*CEFE UMR 5175, CNRS – Université de Montpellier - Université Paul-Valéry Montpellier -EPHE, laboratoire Biogéographie et écologie des vertébrés, 1919 route de Mende, 34293 Montpellier Cedex 5, France.

† Aix Marseille Université, CNRS, IRD, Avignon Université, IMBE UMR 7263, Marseille, France

‡ Aix Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, Evolution biologique modélisation, Marseille, France.

** Equipe ATIP, URMITE UM 63 CNRS 7278 IRD 198 Inserm U1095, IHU Méditerranée Infection, Aix-Marseille Université, Marseille, France

§ XEGEN SAS 15 Rue de la république 13420 Gemenos

Keywords: Conservation biology, Local adaptation, Natural population, Population genomics, Landscape genetics, Landscape genomics

Corresponding author: S. Manel e-mail:stephanie.manel@cefe.cnrs.fr

Running title: Large genomic resources and genome scans

Abstract

Genome scans represent powerful approaches to investigate the action of natural selection on the genetic variation of natural populations and to better understand local adaptation. This is very useful for example in the field of conservation biology and evolutionary biology. Thanks to Next Generation Sequencing, genomic resources are growing exponentially, improving genome scan analyses in non-model species. Thousands of SNPs called using Reduced Representation Sequencing are increasingly used in genome scans. Besides, genomes are also becoming more available, allowing better processing of short-read data, offering physical localisation of variants, and improving haplotype reconstruction and data imputation. Ultimately, genomes are also becoming the raw material for selection inferences. Here, we discuss how the increasing availability of such genomic resources, notably genomes, influences the detection of signals of selection. Mainly, increasing data density and having the information of physical linkage data expand genome scans by i) improving the overall quality of the data; ii) helping the reconstruction of demographic history for the population studied to decrease false positive rates; iii) improving the statistical power of methods to detect the signal of selection. Of particular importance, the availability of a high quality reference genome can improve the detection of the signal of selection by i) allowing matching the potential candidate loci to linked coding regions under selection, ii) rapidly moving the investigation to the gene and function, and iii) ensuring that the highly variable regions in coding regions of the genomes are also investigated. For all those reasons, using reference genomes in analyses of genome scans is highly recommended.

Introduction

Species live in a wide variety of heterogeneous and changing environments, where the environment designs large range of abiotic factors (e.g. temperature, oxygen) and biotic factors (pathogens, symbionts, sexual partners) potentially causing selection. Understanding how these environmental factors lead to genetic adaptation is a longstanding question in evolutionary biology (Kawecki & Ebert 2004). With the rapidly increasing availability of large genomic resources, this topical question can now be addressed by genome scans, i.e. the survey of genetic variability across whole genomes or across a large number of loci in a large number of individuals living in distinct environments (Pritchard 2010). The main objective of genome scans is to identify signals of selection at the genome level in natural populations and thus uncover how natural selection affects genetic variation in response to environmental heterogeneity and changes (Mitchell-Olds *et al.* 2007; Nosil *et al.* 2009; Schluter & Conte 2009).

The identification of these signatures of natural selection in genome scans has become an area of intense research, stimulated by the increasing ease with which a high number of genetic markers can be discovered and characterized in non-model species thanks to Next Generation Sequencing (NGS) (Davey *et al.* 2011). Loci identified as targets of natural selection are likely adaptively and / or functionally important (Vitti *et al.* 2013), and hence candidates for involvement in **local adaptation**[1], disease susceptibility, resistance to pathogens, and other traits of interest to plant and animal breeders. Here, we illustrate our discussion mainly in the field of conservation and evolutionary biology (Allendorf *et al.* 2010; Segelbacher *et al.* 2010). We focus on studies based on the analysis of the spatial sampling of geo-referenced individuals or populations in sites characterized by different environmental conditions. In

[1] Words in bold are defined in the glossary

some cases, the measure of the geo-referenced environmental variable is also available. Those studies aims to better understand local adaptation.

One application of genomes scans in conservation biology is the improvement of the definition of **management units** through integrating adaptive genetic variability (Funk *et al.* 2012). This is particularly important in the case of cryptic genetic lineages which are identified only when using a sufficient number of markers (e.g. Bourret *et al.* 2013; Bradbury *et al.* 2013; Hemmer-Hansen *et al.* 2014; Russello *et al.* 2012). For instance, Bradbury et al. (2013) looked for genomic islands of adaptive divergence (ie. genomic regions associated with adaptation) to refine the delineation of management units in Atlantic cod *Gadus morhua*. They applied an outlier based detection approach (i.e. BAYESCAN, Foll & Gaggiotti (2008)) to 1536 SNPs genotyped for 466 individuals. They detected elevated divergence in 5.2% of SNPs consistent with divergent selection. Those outliers revealed a fine scale geographic differentiation both in the eastern and the western Atlantic, which enhanced individual assignment to the region of origin in comparison to neutral markers. More generally neutral markers **hitchhiking** with selected loci can provide information not only on **positive selection** but also on restrictions to gene flow, which can be used as a tool to identify management units (Gagnaire *et al.* 2015). Another important application of genome scans in conservation biology is the study of species ability to adapt in response to climate change (Pauls *et al.* 2013). The identification of current adaptive polymorphisms can help to identify and maintain adaptive genetic potential. For example Dixon et al. (2015) have identified genomic regions potentially involved in the response to selection for thermal tolerance in the coral *Acropora millepora*. This gives some clues on the possibility of adaptive evolution in future environments, which is essential for conservation (Allendorf *et al.* 2010). Then, detecting climatic variables and polymorphisms potentially involved in local adaptation can improve models of future species dynamics under climate change (Manel *et al.* 2012).

Gaining information on the genetic basis of adaptation to different environments could also be used to improve assisted evolution of natural populations (van Oppen *et al.* 2015).

Here, we focus on genomic data and discuss only the use of NGS–derived markers. **Single Nucleotide Polymorphism** (SNP) data are the marker of choice to conduct genome scans. The physical position of these SNPs on the genome can be known or not, with different degree of precision. SNPs of unknown genomic position are now widely used in non-model species since the recent development of various protocols of **Reduced Representation Sequencing** promoted by the decreasing cost of high throughput sequencing (Davey *et al.* 2013; Davey *et al.* 2011). Alternatively, the recent increase of availability of genomes for a wide variety of species opens the perspective to locate SNPs on the genome using reference genomes (Box 1). The main benefit of moving from thousands of anonymous unpositioned markers with no reference genome to the case of having a reference genome (or long –read sequences), independently of its quality, is to be able to locate and order markers (SNPs) along the genome (Bragg *et al.* 2015) (Figure 1). This is important since in most cases, candidate SNPs identified through genome scans are not the direct target of selection but are rather physically linked to these targets. Then, the knowledge of the **physical linkage** allows the identification of genomic regions, and not only SNPs, with exceptionally high population differentiation (Tine *et al.* 2014). It also allows to explore the number and physical extent of such regions (Nosil *et al.* 2009; Roesti *et al.* 2012a), to derive robust demographic data (Huber *et al.* 2014) and to investigate subtle fine spatial structure (Leslie *et al.* 2015) that are necessary for the inference of selection.

In this review, we discuss the benefits of using additional genomic resources such as reference genomes on the ability of genome scans to detect selection in conservation and evolutionary

biology (Figure 1). We will first discuss which inferences can be made without reference genomes which is still the rule for most species (Gagnaire *et al.* 2012), and then the perspectives that are open by the availability of a reference genome or other genomic resources (Box 1). We then discuss the main challenges remaining even when reference genome is available. First, when genome scans rely on reference genomes, only high quality assembly makes it possible to study the highly variable regions of the genomes in coding regions (Box 2). Second, genome scans only represent a screening tool to reveal potential differences among environments in the functions and histories of the candidate genes identified in these analyses. We therefore discuss more extended validations of the adaptive function of the candidate loci (Pavey *et al.* 2012).

Genomic data used in genome scans

We describe here the diversity of genomic data used to conduct a genome scans in function of the availability of one or several related genomes.

On the one hand, Reduced Representation Sequencing **(RRS)** techniques are based on the use of restriction enzyme digestion to reduce the complexity of the genome and can be performed without prior genomic resources like a reference genome, a transcriptome, or a SNP chip. RRS allow the sequencing of a high number of short genomic regions in different individuals. They notably include Reduced Representation Libraries (RRL), Genotyping By Sequencing (GBS) and restriction-site-associated DNA sequencing (RADseq) (Box 3) and other (RAD, ddRAD, EzRAD, 2B-RAD, nextRAD**)** (Davey *et al.* 2011 ; Peterson *et al.* 2012). Such methods have been increasingly used to produce tremendous amounts of short-reads data to ultimately genotype a high number of SNPs randomly located in the genome of fairly large numbers of individuals. Note that the number of individuals can be increased by multiplexing (at the detriment of individual coverage or the number of sites targeted) or via sequencing pools of individuals (Combosch & Vollmer 2015). Similarly to RADseq, short-

reads sequencing of RNA, RNAseq is also increasingly used to screen polymorphism in DNA coding regions (Piskol *et al*. 2013). Such methods also have huge potential for searching signals of local adaptation since they target cDNA and give at once both DNA polymorphism data and differential expression levels (De Wit & Palumbi 2013). However, the use of transcriptome data can be limited by allele specific expression. Paralogous loci can also induce difficulties in the analysis of transcriptome sequences by creating spurious polymorphic position (Gayral *et al.* 2013). Dedicated approaches are required to limit this biases such as a search of orthologous genes (Li *et al.* 2003 ; Pratlong *et al.* 2015) which can be complemented by the filtering method of SNPs proposed by Gayral *et al.* (2013).

On the other hand, the availability of reference genomes opens different possibilities for generating more powerful individual data to be used in genome scans. First, it is possible to use reference genomes to improve RRS data. The alignment of short-reads produced by RRS methods on a reference genome can notably improve the overall quality of the SNP calling since erroneous reads would be pruned out during the alignment. Physically indexing loci also benefits to data filtering, allowing computing linkage disequilibrium (LD) among physically close loci. Loci in high LD can then be filtered out to allow the use of the dataset in software requiring unlinked loci. Instead of pruning linked loci, the LD data obtained can also be used as a complementary input to limit biases associated with the used of linked loci. Physically locating loci also increases the strength of the datasets since it allows applying a wide variety of selection detection tools, taking advantage of both allele frequencies and loci physical positions (eg F_{ST} sliding windows). Lastly, the distance and LD among loci can also be used during the statistical reconstruction of **haplotypes** (Browning & Browning 2011; Stephens & Scheet 2005).

Moreover, as discussed by Buerkle et al. (2011), analyses of multi-allelic DNA sequences (as haplotypes) rather than bi-allelic markers should help to solve the "n=1 constraint" related to the use of SNP. Besides, haplotypes can be used to improve the imputation of missing data (Browning & Browning 2007). Indeed, if some alleles are inherited more frequently together (haplotypes), and that this information is known, it is crucial to impute data considering these associations rather than only the population frequencies of the alleles being imputed. However, haplotype phasing will be increasingly difficult with increasing recombination rate and decreasing SNP density, which would be an important limitation in studies implementing RRS using relatively rare cutters and single-end short sequencing since sequenced SNP may often fall too far apart.

Genome sequences are also increasingly considered as raw material for genome scans, allowing increasing the number of SNPs detected, to construct haplotypes and at larger divergence scale to perform comparative genomics. Indeed, De novo sequencing can be used to generate a high number of SNPs for multiple individuals. While sequencing parts of the genome using RRS (described above) can bring informative results on adaptive diversity, the number of loci identified through these approaches is also much lower than with genome sequencing (up to tens of thousands compared to millions of SNPs respectively; Rellstab *et al.* 2015). **De novo assembled genome** by deep-sequencing (Chaisson *et al.* 2015a) is one of the most informative approaches for selection detection. However, such data cannot be derived yet for a high number of individuals in conservation projects because it is still too expensive. Nevertheless, a two-step strategy can be used with first de novo assembly of a reference genome by deep-sequencing one individual and followed by the resequencing of additional individuals at lower coverage (Soria-Carrasco *et al.* 2014). In their study of speciation and adaptation to different host plants in the stick insect *Timema cristinae*, Soria-Carrasco *et al.* (2014) first sequenced the genome on the basis of libraries of lengths varying from 170 to

5000 bp. The total genome size was estimated around 1.3 Gb and the assembled genome was estimated to cover around 80% of the genome. Then, Soria-Carrasco *et al*. (2014) re-sequenced at lower coverage the genomes of 160 individuals on eight lanes of Illumina Hi-Seq 2000. The reads were mapped to the reference genome. They thus obtained more than 12.10^6 SNPs with a mean coverage of 5 per SNP and per individual. Such sequencing of several individuals then allowed to produce 4,391,556 SNPs that mapped to one of the 13 linkage group and were used in phylogenetic analysis and to characterize the distribution of genomic variation across the 160 wild-caught *T. cristinae*. However, this approach is still relatively expensive for conservation projects needing numerous individuals. As the power of genome scan approaches can be increased by the number of locations sampled and the number of individuals sampled per location (Lotterhos & Whitlock 2015), other approaches can be useful to reduce sequencing costs. Whole genome sequencing of pooled individuals is a promising and cost effective alternative to individual sequencing (Schloetterer *et al.* 2014). The cost effectiveness of pooled genome sequencing (Pool-Seq) comes from the cost of the preparation of libraries and on the sequencing of more individuals which can increase the estimates of allele frequencies (Schloetterer *et al.* 2014). In their analysis of the pros and cons of Pool-Seq, Schlöterrer *et al.* (2014), indicate their interest to study genetic differentiation, heterozygosity, or selective sweep (Boitard *et al.* 2012). For instance, Turner *et al*. (2010) have sequenced *Arabidopsis lyrata* from four populations (two from serpentine soils and two from granitic soils), with 25 individuals pooled per population, with a 39-fold genome coverage. On the basis of allele frequency differences and F_{ST}, they identified candidate SNPs for the adaptation to soil types. However, Pool-Seq can be limiting for example if one is interested in linkage disequilibrium (Schloetterer *et al.* 2014).

Finally, only high quality genomes will allow studying of highly variable regions located in coding regions (Chaisson *et al.* 2015a) (Box 2). In humans, such plastic regions of the genome are known to contain important immune gene families such as the *Major Histocompatibility Complex* (*MHC*) *class I* and *class II* genes and the *Killer-cell Ig-like Receptor* (*KIR*) genes that are both critical for the resistance to pathogens and hence for the adaptation to different environments (Sommer 2005; Vilches & Parham 2002).

In summary, the above data can be directly used as raw material in the statistical analysis of genomes scans: SNPs (from a small to a high number) or haplotypes.

Detecting the signal of selection without and with a reference genome

Here we discuss how additional genomic resources improve the inferences of the signal of selection in genome scan analyses. We introduce briefly the main principle of the statistical methods, but we do not enter into the details of the available methods and software since multiple reviews have already been published on the subject (Manel *et al.* 2010; Rellstab *et al.* 2015; Schoville *et al.* 2012).

Without any reference genome, SNPs sampled in multiple populations are analyzed as independent variables with methods derived from the test of neutrality of Lewontin and Krakauer (1973) and looking for outlier loci (i.e. loci with higher or lower levels of divergence than expected under neutrality). Loci with high values of F_{ST}, a measure of genetic differentiation among sampled populations, correspond to loci potentially under divergent local selection while loci with low F_{ST} values correspond to loci under balancing selection (Beaumont & Balding 2004).

When environmental variables are available, it is possible to use genetic environment associations (GEA tests, Hedrick *et al.* 1976; Lotterhos & Whitlock 2015) either in a second

round after outlier detection methods or directly for detecting outlier loci. The first application of GEA in the context of genome scan tests the correlation between allele frequencies and environmental variables (Joost *et al.* 2007 ; Schoville *et al.* 2012); then other regression based approaches have been used (Rellstab *et al.* 2015).

However the signature of adaptive processes are not always distinguishable from the neutral genomic background as assumed to have uniform effects across the entire genome, generating a high number of false positive outliers (Schoville *et al.* 2012). Population structure causes correlated allele frequencies and increases the number of false positives generated by genome scans for **selective sweeps** (Excoffier *et al.* 2009). Demographic history can create patterns resembling selection, as in cases of severe bottlenecks, allele surfing during population expansion, secondary contact, and isolation by distance (Novembre & Di Rienzo 2009). Genetic incompatibilities or background selection on gene-coding regions can also confound the signal of selection (Bierne *et al.* 2013). Although, refinements have been introduced to account for those confounding effects (Lotterhos & Whitlock 2015), false positives remain.

Without additional genomic resources, those analyses produce a list of loci that are potentially under selection or linked to alleles under selection, and when available the relevant environmental variables acting as pressures of selection. They generally detect single loci with important effects on adaptation. However some questions may remain, the most evident caveats being that the loci identified as potential targets of selection are usually only statistically linked with close targets of adaptive significance.

We describe below how the reference genome may improve the power of genome scans and help to decrease the number of false positives either through the use of reference genomes to improve previous analysis, or in a direct analysis of raw data.

Mapping SNPs (e.g. identified through RRS loci) to reference genomes can give insights into the potential genes involved in adaptation (see next section) and on the distribution of population genetic statistics along the genome since additional inferences as genetic differentiation and genetic diversity on chromosomal regions become possible with a dense, ordered set of genome-wide markers (Figure 1). In practice, a sliding window sweeps across the genome, to look for regions showing significant deviation of the statistics inferred. Sliding windows are implemented to average locally those statistics and to compare them toward a random set of windows that can be taken from a larger genomic region (eg. the considered chromosome or the whole genome). This sliding-based windows method reinforces the statistical power of the detection of selection since it can highlight regions that may potentially not be spotted using non genomic outlier SNP approaches. It also allows inferring the physical extent of the signal of selection. The size of the window on which the average is calculated and the size of the step are usually defined around 150kb and 50kb, respectively, the step-size being equal or smaller than the windows-size. These windows and step sizes are crucial parameters that can largely influence the detection power of such analyses and should be defined according to the SNP density and sampling variance (Hohenlohe *et al.* 2010). It should be noted that SNP densities obtained in RRS using rare cutters would probably be too small to allow applying with robustness sliding windows along the entire genome.

Comparing the classical local averaged F_{ST} among populations to windows-based F_{ST} would reveal the presence of putative targets of positive selection. For example, this strategy has been used to study the variation of F_{ST} in different lineages of the European sea bass

295 *Dicentrarchus labrax* using a window size of 150 kb (Tine *et al.* 2014) (Box 4). This
296 technique can also be used on whole genome resequencing data, as illustrated in Kardos *et al.*
297 (2015) in which they implemented windows of 100 kb, with step size of 50 kb, and found
298 elevated F_{ST} in regions spanning growth hormone receptor gene. To go even further, Roesti *et*
299 *al*. (Roesti *et al.* 2012b) suggested correcting SNP F_{ST} by local recombination rates. In fact,
300 they showed patterns of within-chromosome large-scale variations in recombination and
301 hitchhiking and thus in F_{ST} that were probably largely influenced by chromosomes structure.
302 They propose to correct for this potential bias by applying sliding window tests on corrected
303 F_{ST} rather than raw F_{ST}, to avoid potential false positive and false negative outliers. Another
304 way of considering this potential bias is to estimate locally averaged recombination rate,
305 nucleotide diversity, and genetic differentiation and to interpret these indices in an integrated
306 manner as illustrated in the European sea bass *Dicentrarchus labrax* (Tine *et al.* 2014).

307 The information of physical linkage among SNPs also opens the possibility to detect reduced
308 genetic diversity in the neighborhood of a selected (directional selection) site along a
309 chromosome, which can be the consequence of **selective sweep** (Nielsen 2005). The extent of
310 the area displaying reduced diversity will depend on the intensity of selection, on
311 recombination rate, on the breeding system and on the age of the selective event. The
312 elimination of slightly deleterious mutations (background selection) can be difficult to
313 separate from selective sweeps (Nielsen, 2005). Specific statistical methods have been
314 developed to detect selective sweeps at the population level (Nielsen 2005; Sabeti *et al.*
315 2006). They differ in the type of the genomic signature of selection detected (Oleksyk *et al.*
316 2010): local reduction in genetic variation at the proximity of the loci under selection
317 (Oleksyk *et al.* 2008), changes in the shape of the frequency distribution of genetic variation,
318 i.e. **site frequency spectrum** (Tajima 1989), extended linkage disequilibrium segments
319 (Sabeti *et al.* 2002; Vatsiou *et al.* 2015), or elevated admixture contribution from one

population (Tang *et al.* 2007). Suh methods relay on the analysis of haplotype. They open perspectives to capture the signal of selection left by both hard and soft sweep. If adaptation has been shaped by new mutations rapidly driven to high frequency in a new environment, those alleles will be detected on haplotypes with low genetic diversity and not observed in ancestral populations or environment (Stapley *et al.* 2010) . In this current issue, Vatsiou *et al* (2015) compared seven methods that either focus on patterns of long range haplotype homozygosity (Sabeti *et al.* 2002) or on the effect of linkage on multilocus genetic differentiation (Chen & Slatkin 2013). Those methods open perspectives to capture the signal of selection left by both hard and soft sweep.

As previously mentioned above, neutral processes can confound the signal of selection. Sophisticated methods using physically indexed SNPs, haplotype or sequence data have been recently developed to improve the inference of demographic processes by helping to reconstruct an accurate demographic history, potentially implicating heterogeneous gene flow along the genome, for the population studied and hence decrease confounding effects (Excoffier *et al.* 2013; Gutenkunst *et al.* 2009; Liu & Fu 2015).

Lastly, the fact that adaptation may rely on polygenic traits with small changes in allele frequencies has been discussed (de Villemereuil *et al.* 2014). One interesting development here is to test outliers not on a single marker basis but on set of genes linked in biological pathways (Daub *et al.* 2013). This is possible only for species with good genomic knowledge, as it requires to assign SNPs to genes on the basis of their physical location and to use functional relationships between genes.

From prioritization of candidate loci to validation

Genome scans characterize loci potentially involved in the adaptation to particular environments but because these methods can generate false positives, a prioritization of candidate loci help to focus on the most reliable candidates. Mapping loci on a reference genome can allow the identification of coding candidates or to loci in physically proximity to coding regions (in this case, a transcriptome can be enough to identify expressed loci). If these genes are annotated in the studied species or a closely related one, it can highlight candidate genes functionally consistent with the studied selective context. For example, using exome sequencing, Yi *et al* (2010) studied the adaptation of human to high altitudes and found that the candidate gene with the strongest signal of selection was a transcription factor involved in response to hypoxia. In this case, the consistence between the function and the environmental context of the study provided a strong support for this candidate. For the non-model organisms for which the reference genome sequence lacks functional annotations, Gene Ontology (GO) could be a useful tool to find pertinent functions in the studied context among the candidate genes (Gu *et al.* 2009). The identification of loci repeatedly involved in the same biological or ecological function in different taxonomic groups exposed to comparable pressures can also be a good argument for the choice of candidate loci (Pratlong *et al.* 2015). For example, Fischer *et al.* (2013) studied the local adaptation of *Arabidopsis halleri* to local climatic conditions by the use of SNPs obtained from Pool Seq. First they detected outlier SNPs on the basis of F_{ST} and differentiation tests. Then they tested the association between these outliers and environmental conditions while controlling for population structure. A GO enrichment was applied to the outlier loci as well and led to the discovery of the implication of functions linked with the response to biotic factors (such as “defence response to bacterium”). Conversely they considered genes with GO potentially involved in the adaptation to climate conditions: among these candidate genes, four were indeed associated with the corresponding environmental factor. This study then underlines how multiple

369 approaches can help identify the factors involved in local adaptation.

370

371 Finally, to improve our understanding of the genomic architecture of adaptation (and
372 speciation) and the nature of the genes involved in these processes, genome scans can be
373 combined with QTL mapping, genome wide association studies (GWAS) and/or sequence and
374 functional analysis (Gagnaire *et al.* 2013; Strasburg *et al.* 2012). In cases of divergent
375 selection, the challenge is to document how functional polymorphisms at individual genes
376 translate into different phenotypes, including quantitative traits values, which in turn translate
377 into fitness differences (Storz & Wheat 2010). In such investigations, the statistical tests used
378 to detect the signatures of natural selection, as well as all the steps before the functional
379 validation, provide the basis for functional inferences: association studies for example link
380 polymorphism at one locus to a phenotypic difference, as was done for the shell color
381 polymorphism in the snail *Cepaea nemoralis* (Richards *et al.* 2013). An important argument
382 to confirm the adaptive function of a particular allele is the design of appropriate experiments
383 to study the resulting effects on protein function or protein expression (Storz & Wheat 2010).
384 This can be done through functional genetic approaches: in the Darwin finches for example, a
385 genomic region with high differentiation levels corresponded to genes previously identified as
386 involved in beak morphology (Lamichhaney *et al.* 2015). For polygenic traits, QTL
387 approaches or GWAS are useful. RAD sequencing was also used to identify QTL associated
388 with ecologically important traits and to gene expression differences in the lake whitefish
389 (*Coregonus clupeaformis*) (Gagnaire *et al.* 2013). The opportunity of whole genome
390 sequencing allows nowadays to move toward GWAS, mostly in the identification of sequence
391 variants associated with risks of complex traits in human genetics, where genomes are
392 available at population scales (Hawley *et al.* 2014). In this field, GWAS was used in the
393 identification of loci linked with the specialized diet of greenlandic Inuit, and associated with

metabolic and anthropometric phenotypes (Fumagalli *et al.* 2015). Nevertheless, the last and probably the most difficult step will be to link phenotypic or genetic differences with fitness differences. Here the analysis of the fitness of all possible point mutants in different controlled environmental conditions open perspectives to make such link. For example, Heiptas *et al*. (2011) analyzed the effect on fitness (growth rate during binary competition in this case) of all possible point mutations for a nine-amino acid region of Hsp90 in yeast. They observed a large proportion of mutation with strong deleterious effect eliminated via purifying selection and few mutation with neutral effect which is consistent with the neutral model of molecular evolution. Linking genotype to fitness open new perspectives such as the possibility to study the population of origin (central population vs margin population) of the mutation in the species range (Rolland *et al. in press*).

Finally as adaptation points to differences in fitness, an experimental validation by comparing different genotypes seems to be the most informative validation step. Experimental tests of local adaptation can be done for example with common garden or reciprocal transplant experiments with different genotypes for candidate loci (Kawecki & Ebert 2004). This should be easier for traits that not relying on too many loci.

Conclusion

In summary, the availability of a reference genome can incontestably improve the detection of the signal of selection at all steps of the genome scans: production of genomic data, statistical detection of the signal of selection, and prioritization of candidate genes. Main improvements can be summarized as follows: i) improving the overall quality of the data; ii) helping to reconstruct an accurate demographic history for the population studied and hence decrease confounding effects (Huber *et al.* 2014) ; with deep coverage, genome sequence data

418 from a single individual are also sufficient to make demographic inference (Miller *et al.* 2012;
419 Oleksyk *et al.* 2012); iii) improving the statistical power of methods to detect the signal of
420 selection. Finally, if the available genome is of high quality, i) it allows matching the
421 potential candidate loci to linked coding regions under selection , ii) it rapidly moves the
422 investigation to the gene and function (Lamichhaney *et al.* 2015), iii) it ensures that the
423 highly variable regions of the genomes in coding regions are also investigated. This last point
424 is important when region under selection are located in such variable regions (Box 2).

425

426 We state that having reference genomes represent a critical step to fully address the question
427 of local adaptation. The community of researchers working in ecology on non-model species
428 will thus gain to collaborate with the community of researchers working on species for which
429 large genomic resources have long been available (human and other model species), as the
430 latter have a long tradition of sequence analysis to detect signatures of natural selection.
431 Combining evolution, ecology, genomics and comparative genomics will help to better
432 address the question of local adaptation, which is topical in a context of global change. Finally
433 we should keep in mind that adaptation *sensu lato* does not only rely on genetic diversity.
434 Indeed the question of non-genetic effects, including epigenetic mechanisms, is of great
435 interest in evolutionary biology (Danchin 2013); yet, here again the use of genomic
436 information should be useful to study the diversity in epigenetic modifications.

437

438

439 **Acknowledgement**

440 We thank the foundation ECCOREV that financed the meetings for this paper. We thank 2
441 anonymous reviewers for very helpful comments. We thank Amanda Xuereb for reading the
442 manuscript. This work is a contribution to the project SEACONNECT funded by the Total

Foundation and to the Labex OT-Med (no. ANR- 11-LABX-0061) funded by the French Government 'Investissements d'Avenir' program of the French National Research Agency (ANR) through the A*MIDEX project (no. ANR-11- IDEX-0001-02). This project has been funded by the ADACNI program of the French National Research Agency (ANR) (project no. ANR-12-ADAP-0016; http://adacni.imbe.fr) and

References

Abi-Rached L, Moesta AK, Rajalingam R, Guethlein LA, Parham P (2010) Human-Specific Evolution and Adaptation Led to Major Qualitative Differences in the Variable Receptors of Human and Chimpanzee Natural Killer Cells. *Plos Genetics* **6**.

Alkan C, Coe BP, Eichler EE (2011) Applications of next-generation sequencing. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363-375.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**.

Baxter SW, Davey JW, Johnston JS*, et al.* (2011) Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. *Plos One* **6**.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Bierne N, Roze D, Welch J (2013) Pervasive selection or is it? why are FST outliers sometimes so frequent? *Mol Ecol* **22**, 2061-2064.

Boitard S, Schloetterer C, Nolte V, Pandey RV, Futschik A (2012) Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. *Molecular Biology and Evolution* **29**, 2177-2186.

Bourret V, Dionne M, Kent MP, Lien S, Bernatchez L (2013) Landscape genomcis in Atlantique salmon (*Salmo salar*) : searching for gene-environment interactions driving local adaptation *Evolution* **67**, 3469-3487.

Bradbury IR, Hubert S, Higgins B*, et al.* (2013) Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications* **6**, 450-461.

Bradnam KR, Fass JN, Alexandrov A*, et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**, 10-10.

Bragg JG, Supple MA, Andrew RL, Borevitz JO (2015) Genomic variation across landscapes: insights and applications. *The New phytologist* **207**, 953-967.

Browning BL, Browning SR (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *American Journal of Human Genetics* **84**, 210-223.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**, 1084-1097.

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**, 703-714.

Buerkle CA, Gompert Z, Parchman TL (2011) The n=1 constraint in population genomics. *Molecular Ecology* **20**, 1575-1581.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Chaisson MJP, Wilson RK, Eichler EE (2015a) Application of Next-Generation sequencing genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics* **16**, 627-640.

Chaisson MJP, Huddleston J, Dennis MY, *et al.* (2015b) Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611.

Chen H, Slatkin M (2013) Inferring Selection Intensity and Allele Age from Multilocus Haplotype Structure. *G3-Genes Genomes Genetics* **3**, 1429-1442.

Church DM, Schneider VA, Graves T, *et al.* (2011) Modernizing Reference Genome Assemblies. *PLoS Biol* **9**, e1001091.

Combosch DJ, Vollmer SV (2015) Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific *Pocillopora* corals. *Molecular Phylogenetics and Evolution* **88**, 154-162.

Danchin E (2013) Avatars of information: towards an inclusive evolutionary synthesis. *Trends In Ecology & Evolution* **28**, 351-358.

Darwin C, Wallace A (1858) On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Proceedings of Linnean Society of London* **45**, 1-8.

Daub JT, Hofer T, Cutivet E, *et al.* (2013) Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution* **30**, 1544-1558.

Davey JW, Cezard T, Fuentes-Utrilla P, *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology* **22**, 3151-3164.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

de Villemereuil P, Frichot E, Bazin E, Francois O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* **23**, 2006-2019.

De Wit P, Palumbi SR (2013) Transcriptome-wide polymorphisms of red abalone (*Haliotis rufescens*) reveal patterns of gene flow and local adaptation. *Molecular Ecology* **22**, 2884-2897.

Dixon GB, Davies SW, Aglyamova GA, *et al.* (2015) Genomic determinants of coral heat tolerance across latitudes. *Science* **348**, 1460-1462.

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844-1849.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *Plos Genetics* **9**.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Fischer MC, Rellstab C, Tedder A, *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology* **22**, 5594-5607.

Foll M, Gaggiotti O (2008) A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993.

Fumagalli M, Moltke I, Grarup N, *et al.* (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343-1347.

Fumagalli M, Sironi M, Pozzoli U, *et al.* (2011) Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *Plos Genetics* **7**.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends In Ecology & Evolution* **27**, 489-496.

Gagnaire P-A, Broquet T, Aurelle D*, et al.* (2015) Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications* **8**, 769-786.

Gagnaire P-A, Normandeau E, Cote C, Hansen MM, Bernatchez L (2012) The Genetic Consequences of Spatially Varying Selection in the Panmictic American Eel (*Anguilla rostrata*). *Genetics* **190**, 725-U703.

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L (2013) The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by rad sequencing *Evolution* **67**, 2483-2497.

Gayral P, Melo-Ferreira J, Glemin S*, et al.* (2013) Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate-Invertebrate Gap. *Plos Genetics* **9**.

Gu J, Orr N, Park SD*, et al.* (2009) A Genome Scan for Positive Selection in Thoroughbred Horses. *Plos One* **4**.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *Plos Genetics* **5**.

Hawley NL, Minster RL, Weeks DE*, et al.* (2014) Prevalence of Adiposity and Associated Cardiometabolic Risk Factors in the Samoan Genome-Wide Association Study. *American Journal of Human Biology* **26**, 491-501.

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environements *Annual Review of Ecology And Systematics* **7**, 1-32.

566 Hemmer-Hansen J, Therkildsen NO, Meldrup D, Nielsen EE (2014) Conserving marine
567 biodiversity: insights from life-history trait candidate genes in Atlantic cod (*Gadus*
568 *morhua*). *Conservation Genetics* **15**, 213-228.
569 Herrera S, Reyes-Herrera PH, Shank TM (2014) *Genome-wide predictability of restriction*
570 *sites across the eukaryotic tree of life*. *bioRxiv*.
571
572 Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape.
573 *Proceedings of the National Academy of Sciences of the United States of America* **108**,
574 7896-7901.
575 Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population Genomics of Parallel
576 Adaptation in Threespine Stickleback using Sequenced RAD Tags. *Plos Genetics* **6**.
577 Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping It Local: Evidence for
578 Positive Selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution*
579 **31**, 3026-3039.
580 Huddleston J, Ranade S, Malig M*, et al.* (2014) Reconstructing complex regions of genomes
581 using long-read sequencing technology. *Genome Research* **24**, 688-696.
582 Iliadis A, Watkinson J, Anastassiou D, Wang X (2010) A haplotype inference algorithm for
583 trios based on deterministic sampling. *Bmc Genetics* **11**.
584 Joost S, Bonin A, Bruford MW*, et al.* (2007) A spatial analysis method (SAM) to detect
585 candidate loci for selection: towards a landscape genomics approach to adaptation.
586 *Molecular Ecology* **16**, 3955-3969.
587 Kajitani R, Toshimoto K, Noguchi H*, et al.* (2014) Efficient de novo assembly of highly
588 heterozygous genomes from whole-genome shotgun short reads. *Genome Research*
589 **24**, 1384-1395.

Kardos M, Luikart G, Bunch R, *et al.* (2015) Whole genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. *Molecular Ecology*, n/a-n/a.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters* **7**, 1225-1241.

Kent WJ, Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Research* **11**, 1541-1548.

Kidd JM, Sampas N, Antonacci F, *et al.* (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods* **7**, 365-U347.

Lamichhaney S, Berglund J, Almen MS, *et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**,

371–375.

Lemaire C, Allegrucci G, Naciri M, *et al.* (2000) Do discrepancies between microsatellite and allozyme variation reveal differential selection between sea and lagoon in the sea bass (*Dicentrarchus labrax*)? *Molecular Ecology* **9**, 457-467.

Lemaire C, Versini JJ, Bonhomme F (2005) Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. *Journal of Evolutionary Biology* **18**, 70-80.

Leslie S, Winney B, Hellenthal G, *et al.* (2015) The fine-scale genetic structure of the British population. *Nature* **519**, 309-314.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* **74**, 175-195.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189.

Liu X, Fu Y-X (2015) Exploring population size changes using SNP frequency spectra. *Nature Genetics* **47**, 555-U172.

Lobreaux S, Manel S, Melodelima C (2014) Development of an Arabis alpina genomic contig sequence data set and application to single nucleotide polymorphisms discovery. *Molecular Ecology Resources* **14**, 411-418.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* **24**, 1031-1046.

Macaulay IC, Voet T (2014) Single Cell Genomics: Advances and Future Perspectives. *PLoS Genet* **10**, e1004126.

Manel S, Gugerli F, Thuiller W*, et al.* (2012) Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology* **21**, 3729-3738.

Manel S, Joost S, Epperson B*, et al.* (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field *Molecular Ecology* **19**, 3760-3772.

Miller W, Schuster SC, Welch AJ*, et al.* (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2382-E2390.

Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics* **8**, 845-856.

Nielsen R (2005) Molecular signatures of natural selection. In: *Annual Review of Genetics*, pp. 197-218.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**, 375-402.

Novembre J, Di Rienzo A (2009) Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics* **10**, 745-755.

639 Oleksyk TK, Pombert J-F, Siu D, *et al.* (2012) A locally funded Puerto Rican parrot
640 (*Amazona vittata*) genome sequencing project increases avian data and advances
641 young researcher education. *GigaScience* **1**, 14-14.
642 Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural
643 selection. *Philosophical Transactions of The Royal Society B-Biological Sciences* **365**,
644 185-205.
645 Oleksyk TK, Zhao K, De La Vega FM, *et al.* (2008) Identifying Selected Regions from
646 Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two
647 Human Populations. *Plos One* **3**.
648 Pauls SU, Nowak C, Bálint M, Pfenninger M (2013) The impact of global climate change on
649 genetic diversity within populations and species. *Molecular Ecology* **22**, 925-946.
650 Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-
651 generation ecological and evolutionary genomics? *Trends In Ecology & Evolution* **27**,
652 673-678.
653 Pavey SA, Gaudin J, Normandeau E, *et al.* (2015) RAD Sequencing Highlights Polygenic
654 Discrimination of Habitat Ecotypes in the Panmictic American Eel. *Current Biology*
655 **25**, 1666-1671.
656 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq:
657 An Inexpensive Method for de novo SNP discovery and genotyping in model and
658 non-model Species. *Plos One* **7**, e37135.
659 Pratlong M, Haguenauer A, Chabrol O, *et al.* (2015) The red coral (*Corallium rubrum*)
660 transcriptome: a new resource for population genetics and local adaptation studies.
661 *Molecular Ecology Resources*, n/a-n/a.
662 Pritchard JK (2010) How we are evolving. *Scientific American* **303**, 41-47.

663 Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RADseq, variant-calling pipeline
664 designed for population genomics of non-model organisms. *Peerj* **2**.
665 Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to
666 environmental association analysis in landscape genomics. *Molecular Ecology*.
667 Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP Allele
668 Frequencies Determined by Pooled Next-Generation Sequencing in Natural
669 Populations of a Non-Model Plant Species. *Plos One* **8**.
670 Rensing SA, Lang D, Zimmer AD, *et al.* (2008) The Physcomitrella genome reveals
671 evolutionary insights into the conquest of land by plants. *Science* **319**, 64-69.
672 Richards PM, Liu MM, Lowe N, *et al.* (2013) RAD-Seq derived markers flank the shell
673 colour and banding loci of the *Cepaea nemoralis* supergene. *Molecular Ecology* **22**,
674 3077-3089.
675 Roesti M, Salzburger W, Berner D (2012a) Uninformative polymorphisms bias genome scans
676 for signatures of selection. *BMC evolutionary Biology* **12**, 94.
677
678 Roesti M, Hendry AP, Salzburger W, Berner D (2012b) Genome divergence during
679 evolutionary diversification as revealed in replicate lake-stream stickleback population
680 pairs. *Molecular Ecology* **21**, 2852-2862.
681 Rolland J, Lavergne S, Manel S Combining niche modelling and landscape genetics to study
682 local adaptation: A novel approach illustrated using alpine plants. *Perspectives in*
683 *Plant Ecology, Evolution and Systematics*.
684 Russello MA, Kirk SL, Frazer KK, Askey PJ (2012) Detection of outlier loci and their utility
685 for fisheries management. *Evolutionary Applications* **5**, 39-52.
686 Sabeti PC, Reich DE, Higgins JM, *et al.* (2002) Detecting recent positive selection in the
687 human genome from haplotype structure. *Nature* **419**, 832-837.

688 Sabeti PC, Schaffner SF, Fry B*, et al.* (2006) Positive natural selection in the human lineage.
689 *Science* **312**, 1614-1620.
690 Schloetterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals-mining
691 genome-wide polymorphism data without big funding. *Nature Reviews Genetics* **15**,
692 749-763.
693 Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National*
694 *Academy of Sciences of the United States of America* **106**, 9955-9962.
695 Schoville S, Bonin A, François O*, et al.* (2012) Adaptive genetic variation on the Landscape:
696 methods and cases. *Annual Review of Ecology Evolution and Systematics* **43**, 23-43.
697 Segelbacher G, Cushman SA, Epperson BK*, et al.* (2010) Applications of landscape genetics
698 in conservation biology: concepts and challenges. *Conservation Genetics* **11**, 375-385.
699 Simpson JT (2014) Exploring genome characteristics and sequence quality without a
700 reference. *Bioinformatics* **30**, 1228-1235.
701 Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology
702 and conservation. *Frontiers in zoology* **2**, 16-16.
703 Soria-Carrasco V, Gompert Z, Comeault AA*, et al.* (2014) Stick Insect Genomes Reveal
704 Natural Selection's Role in Parallel Speciation. *Science* **344**, 738-742.
705 Stapley J, Reger J, Feulner PGD*, et al.* (2010) Adaptation genomics: the next generation.
706 *Trends In Ecology & Evolution* **25**, 705-712.
707 Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype
708 inference and missing-data imputation. *American Journal of Human Genetics* **76**, 449-
709 462.
710 Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer
711 adaptation at specific loci *Evolution* **64**, 2489-2509.

712 Strasburg JL, Sherman NA, Wright KM*, et al.* (2012) What can patterns of differentiation
713 across plant genomes tell us about adaptation and speciation? *Philosophical*
714 *Transactions of the Royal Society B: Biological Sciences* **367**, 364-373.
715 Sudmant PH, Kitzman JO, Antonacci F*, et al.* (2010) Diversity of Human Copy Number
716 Variation and Multicopy Genes. *Science* **330**, 641-646.
717 Tajima F (1989) Statisticial method for testing the neutral mutation hypothesis by DNA
718 polymorphisms. *Genetics* **123**, 585-595.
719 Tang K, Thornton KR, Stoneking M (2007) A New Approach for Using Genome Scans to
720 Detect Recent Positive Selection in the Human Genome. *PLoS Biol* **5**, e171.
721 Tine M, Kuhl H, Gagnaire P-A*, et al.* (2014) European sea bass genome and its variation
722 provide insights into adaptation to euryhalinity and speciation. *Nature*
723 *Communications* **5**, 1-10.
724 Turner JRG (2010) Population resequencing reveals local adaptation of A*rabidopsis lyrata* to
725 serpentine soils. *Nature Genetics* **42**, 260-263.
726 van Oppen MJH, Oliver JK, Putnam HM, Gates RD (2015) Building coral reef resilience
727 through assisted evolution. *Proceedings of the National Academy of Sciences* **112**,
728 2307-2313.
729 Vatsiou AI, Bazin E, Gaggiotti OE (2015) Detection of selective sweeps in structured
730 populations: a comparison of recent methods. *Molecular Ecology*,.
731 Vilches C, Parham P (2002) KIR: Diverse, rapidly evolving receptors of innate and adaptive
732 immunity. *Annual Review of Immunology* **20**, 217-251.
733 Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting Natural Selection in Genomic Data.
734 *Annual Review of Genetics, Vol 47* **47**, 97-120.
735 Wollstein A, Stephan W (2015) Inferring positive selection in humans from genomic data.
736 *Investigative genetics* **6**, 5-5.

Yi X, Liang Y, Huerta-Sanchez E, *et al.* (2010) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329**, 75-78.

Glossary

De novo assembled genome: sequence of a genome from overlapping DNA sequences without having a reference genome.

Haplotype : set of genetic variants or alleles physically linked on a chromosome and statistically associated. They are inherited together until disrupted by recombination.

Hitchhiking (to fixation): when a genetic variant that is physically linked to a selectively advantageous mutation goes to fixation because of the selection on that advantageous mutation.

Linkage disequilibrium (LD): the non-random association of alleles at different loci. LD can strongly increase around selected sites when selection is strong.

Local adaptation: the result of fitness differences between alleles (Wollstein & Stephan 2015). A fitness value can be assigned to each possible genotype.

Management unit: Genetically differentiated population to be managed separately because of its demographic independence

Negative selection (or purifying selection): any type of selection where new mutations are selected against (Nielsen 2005). In this case, the derived allele is detrimental to the organisms.

Physical linkage describes the tendency of alleles near each other on a chromosome to be inherited together

Positive selection as introduced by Darwin and Wallace (1858)), "is the principle that beneficial traits – those that make it more likely that their carriers will survive and reproduce- tend to become more frequent in populations over time" (Sabeti *et al.* 2006). Then, it is any type of selection where coefficient of selection has a positive value. It occurs when the derived allele has a higher fitness than the ancestral allele. It comprises both directional selection and balancing selection as overdominance. Positive selection is the mechanism that can lead to local adaptation (Nosil *et al.* 2009; Pritchard 2010).

Reduced Representation Sequencing (RRS): a technique based on the use of restriction enzyme digestion to reduce the complexity of the genome.

Selective sweeps: reduction or elimination of genetic variation in a genomic region as the result of natural selection having favored one particular variant in this region.

Site frequency spectrum. The genomic signatures of recent adaptations can be measured by the site frequency spectrum (SFS), which summarizes the counts of derived variants in a region.

Soft sweep: Selection can occur either on standing genetic variation, i.e. existing genetic variation, creating a **soft sweep** or on a *de novo* mutation, creating a **hard sweep.**

Legend of figures

Figure 1. Schematic representation of the benefits of using additional genomic resources (mainly reference genomes) in the different steps of genome scans. GBS: Genotyping by sequencing.

Boxes

Box 1- A reference genome

A reference genome is an assembly of a genome that is representative of a species and can be used to align sequence reads for population-genomic studies. In genome scan analyses, a reference genome can either be the genome of the targeted species or that of a closely related species. Similarly, this reference can be based on a single individual or on a collection of individuals; in the former situation however, the reference genomes do not capture the full extent of nucleotide or structural variation segregating within a species. With the dramatic decrease in the cost of DNA sequencing of the past few years, an ever increasing number of genome are characterized, yet the quality of these genomes can be highly dependent upon the quality of the assembly (Simpson 2014). As further developed below, reference genomes can thus be of two types: they can be represented by a single sequence ('golden path' reference genomes) or they can be representative of the genetic diversity of the species ('high quality' reference genomes). For these 'high quality' reference genomes, genetic variation does not just refer to single nucleotide polymorphisms (SNPs) but also to structural variation that represents a large source of genomic diversity (Church *et al.* 2011).

High quality genome (like human, mouse or *A. thaliana*).

A 'high quality' reference genome is defined here as a reference sequence that is both complete in terms of coverage and representative of the genetic diversity of the species investigated. Because current sequencing technologies can only target the euchromatic portion

of the genome (hence excluding the heterochromatin that includes the centromere and telomere regions), 'complete' refers here to the euchromatic genome. 'Finished' quality genome projects typically cover >95% of the euchromatin sequence, a target that can however be difficult to achieve depending on the genome size and complexity of the targeted species. Here genetic variation does not just refer to single nucleotide polymorphisms (SNPs) but also to structural variation that represents a large source of genomic diversity (Church *et al.* 2011).

With the notable exception of the human genome however, most genome sequences only satisfy the first criterion (‘complete’ sequence) and thus represent what was defined as a 'golden path' during the course of the sequencing of the human genome: a non-redundant haploid representation of the genome (Kent & Haussler 2001).

A 'golden path' reference genome

A 'golden path' reference genome thus represents the first level of reference sequence that could be used for genome scans. It is now possible to rapidly generate such a reference sequence thanks to modern sequencing technologies that have led to a rapid decrease in the cost of sequencing: for example in 2014-15, typical cost for re-sequencing of a human-size genome (3,000Mb) with paired-ends reads at 30X coverage on an Illumina HiSeq platform is $4,211. This cost includes labor, administration, management, utilities, reagents, and consumables; sequencing instruments and other large equipment (amortized over three years); informatics activities directly related to sequence production; submission of data to a public database and indirect costs (source: National Human Genome Research Institute (NHGRI); http://www.genome.gov/sequencingcosts/).

With sufficient coverage (>30X), this approach should lead to a reference genome covering >95% of the euchromatin sequences. Those values are however highly dependent on the size of the genome and on factors such as the frequency of repeated regions in the genome. Such complete sequences are rarely obtained in conservation biology at the moment however, as illustrated by the de novo sequencing of the genome of *Arabis alpina*, which resolved less than 50% of the genome (172 Mb out of 370 Mb) (Lobreaux *et al.* 2014**).** Similarly, complex regions with recently duplicated segments for example or plastic regions in diploid genomes may not be well resolved by this approach (Alkan *et al.* 2011).

Improving quality of assembly

Over the past couple of years, ways to improve initial genome assemblies have emerged. In particular, technology from Pacific Biosciences (PacBio) for long-read single molecule, real-time (SMRT) sequencing can upgrade genomes to a higher quality finished state (Huddleston *et al.* 2014). SMRT sequencing is however still costly and can thus only be applied to small genomes or to targeted resequencing of specific regions. To avoid the problems associated with sequencing plastic regions in diploid genomes, an interesting alternative is also to work with haploid cell lines generated from the same organism such as specific tissues Chaisson *et al.* 2015b) or specific life stages (Rensing *et al.* 2008). Although such cell lines will likely not be available for most organisms, single cell genomics are rapidly progressing so that single gametes could become an adequate resource to establish reference genome (Macaulay & Voet 2014). The sequencing of the genome of the European Sea Bass was based on the use of a meiogynogenetic individual (Tine *et al.* 2014). Finally, in addition to the experimental improvements, the quality of the reference genomes also increases with the development of better assemblers, i.e. the software used to assemble genomes. Indeed, software now exists to help *de novo* assembly from whole – genome shotgun short reads (Kajitani *et al.* 2014) and efforts are being made to evaluate the performance of these methods (Bradnam *et al.* 2013).

Status on genome sequencing

Progress on genome information can be found here: http://www.ncbi.nlm.nih.gov/genome/browse/#

Box 2- Challenges of highly variable regions

While progress is being made to generate 'golden path' reference genomes, this will at best lead to the complete sequencing of one haploid genome in a given species. Such a sequence will be sufficient for the analysis of genome scans for the vast majority of genes and genomic regions but will be problematic to investigate the more plastic regions of the genome. Indeed, even for the highest quality mammalian genome, the human genome, it was found that some genomic regions were missing in the reference sequence (Kidd *et al.* 2010). These specific regions display structural differences between the genomes of the individuals investigated and the reference genome, either because of specific deletions or because of underrepresented multi-copy genes in the reference genome (Sudmant *et al.* 2010).

In human such plastic regions of the genome are known to contain important immune gene families such as the *Major Histocompatibility Complex* (*MHC*) *class I* and *class II* genes and the *Killer-cell Ig-like Receptor* (*KIR*) genes that are critical for the resistance to pathogens and hence for the adaptation to different environments (Sommer 2005; Vilches & Parham 2002). Indeed recent genome scans approaches on human populations underlined the critical roles of pathogens in local adaptation (Fumagalli *et al.* 2011). Both gene families display a high level of polymorphism but also gene-content variation (see Figure for the *KIR* locus) so that a single reference sequence will not represent correctly the gene content in all individuals and could lead to problematic analyses for genome scans. Indeed, not including multiple references for such regions produces misalignments of the reads and spurious variant calls. The human reference genome was thus organized to include multiple sequences for these regions (Church *et al.* 2011). Thus 'high quality' reference genomes that include multiple sequences for the plastic regions are necessary to produce unbiased results in genome scans. Such a reference sequence requires two additional steps once a first complete sequence of the genome is known: the identification of these regions and the characterization of the variation at these regions. Finally, data from these regions might need to be adapted to run some of the usual tests of local adaptation such as F_{ST} outliers: for example by encoding presence/absence of genes (*KIR* locus) as SNPs or by considering complete alleles rather than individual SNP (*MHC* genes). Because these highly variable gene families have a strong impact on immune adaptation, they are often investigated by targeted analyses, outside of genome scans (candidate gene approaches)

898

899 **Figure.** Example of a highly-variable region of the genomes of primates: the Killer-cell Ig-
900 like Receptor (KIR) locus. This drawing compares the organization and gene-content
901 variability of the human and chimpanzee *KIR* loci. The branching pathways illustrate how
902 different gene-content motifs can combine to produce different *KIR* haplotypes. Genes that
903 are typically found on all haplotypes of both species (framework genes) are colored grey;
904 chimpanzee-specific KIR are colored green. Humans have two broad groups of haplotypes
905 called A and B haplotypes that differ in gene content and level of allelic variability (Parham *et*
906 *al.* 2012): genes characteristic of human A haplotypes are colored red while genes
907 characteristic of human B haplotypes are colored blue (2DP1 and 2DL1 in humans are
908 colored grey to indicate their presence both on A and B haplotypes). Adapted from Abi-
909 Rached et al. (Abi-Rached *et al.* 2010).

910

911

912

Box 3: Methods based on the use of restriction enzyme digestion of target genomes to reduce the complexity of the target: example of RAD-seq

RAD-Seq is based on the sequencing of short sequences flanking restriction sites (Baird *et al.* 2008). It requires choosing the restriction enzyme(s) in combination with the sequencing depth and the number of individuals multiplexed in order to modulate the number of expected markers. This choice is done on the basis of genome knowledge (size and GC-content) for the target species or on phylogenetically related species as proposed by Herrera *et al.* (2014). The analysis of RAD-Seq data can be done thanks to different software packages such for example as Stacks (Catchen *et al.* 2013), PyRAD (Eaton 2014), RADtools (Baxter *et al.* 2011), GATK (DePristo et al. 2010), or dDozent (Puritz *et al.* 2014).

As an illustration, (Pavey *et al.* 2015) used RAD-Seq to study the genetic basis of the differentiation between ecotypes of the American eel (*Anguilla rostrata*). They used the EcoRI restriction enzyme and analyzed 379 individuals with around 24 individuals per lane of an Illumina HiSeq 2000. After cleaning and filtering the data they retained 42 424 SNPs which corresponded to 1 SNP every 40 kb. This density allowed them to identify 331 SNPs associated with the ecotypes, with 99 SNPs corresponding to annotated protein-coding genes.

Box 4- Genome scan and genomic resources: a case study on the European sea bass (*Dicentrarchus labrax*)

Population genomic studies can be useful both for the management of wild and domestic hatchery populations and for a better understanding of the evolution of species, which is useful for conservation purposes. The study on the European sea bass (*Dicentrarchus labrax*) provides a good illustration of the multiple interests and applications of genomic data. This species is heavily harvested. Its distribution range, from the Atlantic Ocean to the Black Sea, encompasses an important barrier to gene flow at the Almeria-Oran front which separate two genetic lineages (Lemaire *et al.* 2005). The European sea bass can be present in different levels of salinity which raises the question of its adaptation to different environments (e.g (Lemaire *et al.* 2000)). Getting genomic information on this species would be useful to better understand its evolutionary history and would be useful for aquaculture purposes. Tine et al. (2014) published the first draft genome of *D. labrax*. The sequencing of the genome was based on a meiogynogenetic male, with an average coverage depth of 30X and used a combination of whole-genome shotgun, mate pair and BAC end sequencing. The length of the assembled genome reached 650 Mbp and 86% of the contigs were assigned to the 24 chromosomes of this species. The genome annotation led to the identification of 26,719 genes and an important collinearity was observed with other teleosts genomes. This genome corresponds to the 'golden path' genome as described in Box 1. RAD-sequencing was then used to study the polymorphism at the genome scale for 100 individuals (and three from anoutgroup, *D. punctatus*). The authors obtained around 178,000 RAD loci which, after mapping on the genome, allowed them to analyze one locus every 7.5 kb and to reach a 2.5% genome coverage. The combination of the reference genome and RAD loci allowed for a better understanding of the evolutionary and adaptive history of the species. For example, the identification of signatures of positive selection in duplicates of genes involved in osmoregulation (e.g. *PRL-L2*) opened the way to a better understanding of euryhalinity in this species and in other teleosts as well. Here important information comes from the combination of a good annotation, the discovery of multiple gene copies of the same family and from statistical test of molecular evolution. The genomic analysis of differentiation between Atlantic and Mediterranean lineages led to a mean F_{ST} of 0.28 but with highly heterogeneous repartition: genomic islands of differentiation were observed with lengths varying from several hundred kb to more than one Mb. These islands were negatively correlated with local recombination rate and diversity levels (with a few exceptions). Mapping SNPs obtained

967 through RAD sequencing to a reference genome is here useful to understand the evolution of
968 genomic differentiation and ultimately of speciation. Similar approaches could be applied to
969 the study of adaptation to different environments. Having such genomic information,
970 including gene annotation and knowledge on the evolution of gene contents (i.e. duplications)
971 is a highly valuable complement to the analysis of isolated SNPs.

972

973

Steps in genome scans

Benefits

Genome sequences / Genomic ressources

(1) Genomic data

- Hypervariable regions
- Physical linkage
- Haplotypes

High quality genome

SNPs obtained from reduced representation sequencing; High density of SNPs obtained from genome resequencing ;

(2) Detection of the signal of selection

- Sliding windows statistics
- Haplotype analysis: demographic inference ; signal of selection

(3) Prioritization

- Detection of the physical closest coding gene
- Detection of the functions of the gene

High quality genome

Figure 1