Expression of interest for a network of excellence in
POPULATION GENOMICS

Coordinator: Michel Veuille <mveuille@snv.jussieu.fr>
Supplementary information <http://gdrevol.snv.jussieu.fr/pgmics.html>

European priority :
1.1 Genomics and biotechnology for health
1.1.1 Advanced genomics and its applications for health
1.1.1.i.a Fundamental knowledge and basic tools for functional genomics in all organisms
Comparative genomics and population genetics

1. Definition: what is population genomics

Population genomics is the study of the forces that determine patterns of neutral and adaptive variation in genomes. These forces include i) the origin of variation (the rates of mutation and of recombination; the mobilisation of transposable elements in a host genome), ii) the distribution of variants within a species (e.g. the genetic drift of allele frequencies in finite populations; the migration of individuals among populations), and iii) the consequences of molecular changes (the adaptation of organisms to their environment through changes in gene expression; the coevolution or conflict of DNA elements within the genome).
The purpose of the "population genomics" network is to understand these factors and their effects on genome evolution. Recently, the genomes of several organisms have been fully sequenced. We will use this rich source of data as a basis for our studies.
Our network is primarily interested in understanding the causes of molecular variation. However, population genomics lies upstream of an important field of application, based on the principle that DNA carries large amounts of information: in addition to the functional information which it encodes, it also carries historical information. Molecular variants in populations generally have ancient origins (on average they are about a half million years old in humans). The patterns of genetic variation and associations between variants at different genomic locations ("linkage disequilibrium") record the influence of past events. For instance, human genetic variation shows traces of the migration of the human species out of Africa, of its population expansion at the end of the Palaeolithic, of its subsequent structuring into local populations, and of recent admixture among huge populations. Retrieving and interpreting this information is one of the most challenging issues for attempts to characterise disease genes in our own species, and to improve crop and livestock species. Population genomics also has consequences for future programmes of conservation genetics, which have important implications for the management of natural resources.

Another essential aspect of population genomics is the role of comparative studies. Ever since the evolution of the eukaryotic genome characteristic of higher organisms (those with a cell nucleus), the general organisation of the chromosome machinery, its gene content, and the nature of these genes have changed remarkably little. For over a billion years, the eukaryotic genome has been a powerful evolutionary machine. Evolutionary change has occurred by means of a process of turnover of mutations that have greatly modified the sequences and regulation of genes, and by fixation of duplicate copies of genes. Since the basic properties of the genome unite the living world, population genomics as a science is not limited to any single species (e.g. humans). Many research materials can contribute to a general view of genomic evolution. These research materials must include both the human genome, and the genomes of model organisms that are easier to work with like yeast, mice, *Drosophila*, and *Arabidopsis*.

2. Need and relevance

A fundamental field with many downstream applications. This network falls under the "Advanced genomics and its applications for health" priority. It is also important for conservation biology, which pertains to the "sustainable development, global change and ecosystems" priority. In addition to its integrative role for the production of knowledge, this network will also play an important role in transmitting knowledge in an emerging fundamental field. Below, we present four examples that illustrate the importance of this expertise for applied research situated downstream of our network.

Example 1: mapping genes in humans. There are about 30,000 genes in the human genome. An important method for locating complex disease genes in the future will use the neutral single nucleotide polymorphisms (SNPs) that are abundant in populations. They are associated into small groups of variants that co-segregate as "haplotypes" along the DNA sequence. Association studies attempt to localize the SNPs that happen to lie in the neighbourhood of disease genes. A recent review estimates that mapping programmes will require about 300,000 SNPs. Ideally, the density of markers and the strength of their associations should depend only on the size of the species and on the recombination rate. In practice, however, their distribution patterns are affected by several other factors, which our network plans to study: i) the rate of occurrence of deleterious mutations ("background selection"), ii) the rate of spread of advantageous mutations ("selective sweeps"), iii) demographic changes in the history of species. Understanding these factors may prove critical in the successful use of association studies. Both theoretical and experimental advances are needed to handle and interpret this enormous wealth of information in actual populations. For instance, we have no idea of the rate of favourable selection events in the past history of any species. Similarly, the number of deleterious mutations per generation per genome remains unknown except in a handful of species. The purpose of the Population Genomics network is to contribute to understanding the interplay of these factors across the genome.

Example 2: mapping genes in domestic species. The problems are essentially the same as above, except that experiments are possible in non-human species. The mouse and *Arabidopsis* systems are useful genomic models that are biologically close to many domesticated species.

Example 3: conservation genetics. The demographic and industrial pressure upon natural ecosystems leads to the fractionation of the range of many species and to a decrease in their population size. A challenge for ecology is to prevent their extinction. In addition to demographic risks, many geneticists expect a genetic deterioration of endangered species. When populations are confined to a low population size for a long time, mildly deleterious alleles can fix by genetic drift, and the adaptation of populations to their environment can show a long term decrease. This is the "mutational meltdown", a hotly debated issue in evolutionary biology.

Example 4: expression of adaptive traits. Until recently, the genetic study of adaptive traits involved one of two methods. A highly successful approach involves the "genetic dissection" of model organisms using experimentally induced mutations. This method is slow and can only focus on a small number of genes. The other method is the statistical analysis of quantitative trait variation. This can provide a global estimate of the amount of genetic variability in a trait in a given population, but does not tell us the number of genes or the level of variation contributed by individual genes. Results for one population in a given environment cannot necessarily be extrapolated to another population in the same environment or to the same population in another environment. The study of the transcriptome using microarray techniques allows us to proceed downwards from the genes to the trait. In genomic model systems, it is possible to carry out an exhaustive count of the

genes involved in a given function, to study their expression in changing conditions, and to record their variation between individuals.

In summary, our network will occupy a unique place in European priorities in biotechnology. It will provide critical information on factors that affect the genomic origin and evolution of variation in natural populations. This field has potential applications in many other fields, including industrial and medical research. We are not ourselves concerned with these applications at this stage. We will focus on species whose genomes have been completely sequenced, together with their close relatives, since improving knowledge about genome evolution is fundamental to our approach. Other species will be considered when they are more suitable than available genomic systems.

3. Integrative and structuring actions

Our specific activities will include **structuring actions between European laboratories** and the **transmission of knowledge**. Structuring actions will incite European researchers to collaborate. They will be divided into **broad-scale projects** and **meetings**. Broad-scale projects will be campaigns of data collection. Free access to the results will be guaranteed by the consortium. These campaigns will last only for the duration of the 6th framework programme (five years), so that laboratories will be free to quit when it is completed. We plan to conduct two kinds of projects: sequence polymorphism studies, and expression studies. Structuring actions will also develop theory and bioinformatic tools. The transmission of knowledge will involve both training through mobility and a summer school.

(a) Meetings. They are a critical step for people to collaborate. The consortium will organize open workshops.

(b) Sequence polymorphisms. This project will concern yeast, Drosophila, Arabidopsis and humans. For these taxa, the collection of polymorphism data in stratified samples of populations will assess the relative importance of background selection/selective sweeps/demography on levels of segregating variation and of linkage disequilibrium. The advantage of Drosophila lies in its high level of variation (it is more than ten times more polymorphic than the human genome) and the fact that regional variation in its recombination rate is well-characterised. Furthermore, as in humans, *D. melanogaster* and *D. simulans* have undergone a bottleneck of small population size in spreading from Africa to the rest of the world several thousand years ago. We will carry out this project on a substantial part of the Drosophila genome (several megabases). We will collect polymorphism data in the human genome Patterns of polymorphisms in non-coding regions will be contrasted to those at coding or potentially selected gene regions. This contrast is necessary to differentiate molecular signatures of demographic processes from those expected under several forms of selection. This distinction could allow us to better design studies aiming at recognising adaptive and/or disease-associated loci in genome-wide screens, which is a major challenge of future genomics work. Different species of Arabidopsis and yeast differ in their breeding systems, causing different effective levels of genetic recombination. Comparative studies of the distributions of variants within and between populations and species will provide important data relevant to the effect of background selection/selective sweeps/demography that complement those from within-genome comparisons in Drosophila and humans.

(c) Expression studies. These will be carried out using several model organisms: mice, *Arabidopsis*, *Drosophila*. (see example 4 above).

(d) Training through mobility will be essential, however this is not specific to our network. We will also encourage collaboration visits between laboratories.

(d) Summer school. Above we showed that the field of population genomics represents a fundamental source of knowledge for a number of biotechnological applications

including medical genetics, agricultural research and conservation genetics. **The transmission of this knowledge is fundamental for this network**. Theoretical population genomics seems a complex science to most biologists, since it involves a continuous interplay between empirical data and elaborate mathematical models. Even though this "expression of interest" insists on the close proximity between population genomics and applied genomics, we are conscious that a wide gap exists between them. Bridging this gap will be a major contribution of our network to the European Research Area. We plan to organise a summer school on a yearly basis. It will provide teaching to researchers, to postdocs and to graduate students from the network and from outside the network. It will also visit a different country each year, and will provide teaching in the universities visited whenever possible.

4. Excellence

The laboratories participating in the core group played an active part in the rise of population genomics. Below we give a brief account of their contribution.
Since the 1980s *Griffiths* and *Donnelly* (now in Oxford) have participated along with John Kingman, from Bristol, in the rise of the "coalescent theory", a mathematical framework analysing molecular variation from the genealogical nature of DNA evolution. They are still among the leading theoreticians in this field. In 1993 B. and D. *Charlesworth* (now in Edinburgh) put forward (along with Morgan) the "background selection" hypothesis, which predicts that the elimination of recurrent deleterious mutations in the genome exhausts nucleotide polymorphism in regions of low recombination rate. This mechanism appears to be an important factor determining the density of single nucleotide polymorphims in *Drosophila* (Hudson and Kaplan 1996, Charlesworth 1996) and in humans (Nachman 2001). Since 1989 several members of this group who work on *Drosophila* (*Aguadé, Brookfield, Stephan, Schlötterer*, ..) have contributed to the exploration of the hitch-hiking effect (a hypothesis put forward in 1974 by Maynard Smith and Haigh), which makes it possible to detect the signature of natural selection in genomes in populations. This issue has also been examined from a theoretical perspective by *Przeworski, Wiehe,* and *Stephan.* European researchers have provided estimates for the load of deleterious mutations per genome per generation in *Drosophila* (*B. Charlesworth* 1994) and in humans (*Eyre-Walker* and *Keightley* 1999).

5. Policy context and expected benefit for the European research area

Population genomics is a new field of population genetics with strength in both the US and Europe. The attraction of North America for population geneticists around the world is illustrated in our core group: many of our members have worked in American laboratories and established long-term collaborations with them. The leading role of US population genetics dates from the 1970s with the decision of the NIH to fund several population genetics groups after the breakthrough in the study of molecular variation by Lewontin and Hubby in 1966. This linked population genetics to the rising field of molecular biology, traditionally closer to the health domain. In continental Europe population genetics retained stronger associations with ecological studies. Population genetics is, of course, useful in ecology for analysing the population structure, but its focus remains genetical and evolutionary. Britain, however, had an eminent role developing population genetics and also pioneered the rise of molecular population genetics of humans. However, this is a relatively small-sized community that will benefit greatly from integration into a European network.

The resources for an active population genomics community exist in Europe; however, it is essential for Europeans to more actively collaborate with each other. This will increase their productivity and their relations with the other life sciences, at a time when population

genomics is becoming increasingly important for the health sciences. The US community will also benefit from an international partnership with an active European research.

6. Dimension of the project and critical mass

We can gather a workforce that will be visible at the international scale, and provide teaching in population genomics at a high level throughout Europe, including non-member states. We will forge links between laboratories in a number of countries during our FP6 contract. We also hope to bring the importance of this field to the attention of European universities. Note that this is a fundamental research field in which productive collaborations occur with scientific communities in other countries (US, Japan, Australia and Canada). Since this field is relatively small worldwide, an integrated European research network can contribute to bringing the international community up to a substantial working capacity, and can engage in a fruitful synergy with others.

7. Composition of the core group

The core group is comprised of about 170 researchers from 45 research groups in 11 countries. It represents the essential part of the field in Europe. It includes pure theoreticians and geneticists working on genomic models. A detailed presentation of the group is available on our website (http://gdrevol.snv.jussieu.fr/pgmics.html).

Austria	Christian Schlötterer , Reinhard Bürger (Wien): **6 researchers**
Belgium	Xavier Vekemans (Bruxelles): **3 researchers**
Denmark	Mikkel H. Schierup (Aarhus): **4 researchers**
Finland	Pekka Pamilo, Outi Savolainen (Oulu): **6 researchers**
France	Laurent Duret, Christian Biémont, Lluis Quintana-Murci (Lyon-1), Philippe Jarne, Pierre Boursot, Alain Bucheton (Montpellier), Brigitte Crouau-Roy (Toulouse), Michel Veuille (Paris-6), Evelyne Heyer (Paris-7), Marie-Louise Cariou (Gif), Anne Atlan (Rennes), Gordon Luikart (Grenoble): **43 researchers**
Germany	John Parsch, Wolfgang Stephan (München), Molly Przeworski (Leipzig), Thomas Mitchell-Olds (, Jena), Thomas Wiehe (Berlin), Dietard Tautz (Köln): **31 researchers**
Poland	Ryszard Korona (Jagiellonian University): **5 researchers**
Portugal	Jorge Vieira (Porto), Isabel Gordo (Instituto Gulbenkian de Ciencas): **5 researchers**
Spain	Montserrat Aguadé, Jaume Bertranpetit, Alfredo Ruiz (Barcelona): **15 researchers**
Switzerland	Laurent Excoffier (Bern): **4 researchers**
U. K.	Nicholas H. Barton, Brian Charlesworth, Deborah Charlesworth, Peter Keightley (Edinburgh), Mark Beaumont (Reading), John Brookfield (Nottingham), Austin Burt (Imperial College), Laurence D Hurst (Bath), Edward J Louis (Leicester), Peter Donnelly, Robert Griffiths, Jotun Hein, Gillean McVean (Oxford), Adam Eyre-Walker (U. of Sussex) : **44 researchers**

8. Organisation of the network

The network will be headed by a coordinator and a council who will plan projects and prepare the budget on a yearly basis. The main activities will be conducted by specific coordinators. They will concern:

1. Meetings,
2. Genome variation projects
3. Genome expression projects
4. Training and mobility
5. Summer school

Activities 2 and 3 are data creating operations. They will be co-directed by two coordinators, each working on a different model system, in order to avoid specialisation in the network.