Next-generation high-throughput sequencing technologies

Technological worksheet November 2008
Catherine Golstein and Michel Caboche, Technologies of the Future, INRA

1. Introduction, background

Sequencing evolution, achievements and limitations

Published by Sanger and colleagues in 1977, the dideoxynucleotide method for DNA sequencing remained the standard for the next 30 years. Providing a tool to decipher the genetic blueprint of all life on Earth, Sanger sequencing transformed biology. From single genes to whole genomes, from prokaryote to eukaryote genomes, the era of Sanger sequencing culminated in 2001 in a milestone for human history: the completion of the human genome sequence.

Over the years, and driven lately by the human genome objective, Sanger sequencing went over many technical improvements in throughput, accuracy, safety, robustness and sensitivity. Notably, the radioisotope labels of dideoxy-terminator nucleotides were abandoned for base-specific fluorescent dye labels, the slab sequencing gel allowing the electrophoretic separation of sequencing products replaced by a capillary array platform. Sequencing a complex genome with Sanger technology today is estimated to cost about €25 million for several years of intensive work. Although this is a significant progress compared to the cost of the Human Genome Sequence Project (over 10 years and $1 billion), there is a need for further scaling-up throughput and minimising cost of sequencing to meet the growing demands of research in human genetics and genomics, agriculture and environmental science.

Most lingering limitations to further lowering the cost and increasing the throughput of Sanger sequencing are inherent in the technology. With the reference goal of the "$1,000 genome", *i.e.* $1,000 for *de novo* sequencing of a human-size genome, as established by the American National Institute of Health in its 2004 request for proposals, the race is on for novel, cheaper and faster sequencing technologies.

2. What are the next-generation sequencing technologies?

How do they work? What make them new technologies? What do they bring to previous technologies?

The era of Sanger monopoly on sequencing is over. Launched between 2005 and 2007, three technologies of the second- or so-called next-generation sequencing technologies are now competing with Sanger: the **GS (Genome Sequencer) FLX System** from Roche (previously 454 sequencing technology, developed by 454 Life Sciences), the **Illumina Genome Analyzer** (previously Solexa 1G, developed by Solexa), and the **SOLiD (Supported Oligonucleotide Ligation and Detection) DNA Sequencer** from Applied Biosystems**.**

Staple of the next-generation sequencing technologies is their ability to sequence **massive amounts of templates in parallel**, producing in one run hundreds of thousands of reads for GS FLX, tens of millions for Illumina GA and over a hundred million for SOLiD, where the latest Sanger platform produces only 96 reads at a time.

This increase in throughput was accompanied by a **dramatic drop in sequencing cost per base**, down to a fraction of a percent that of Sanger sequencing. In addition, the new technologies have in common to **overcome the bias due to bacterial cloning** in Sanger, and to provide a **quantitative readout** for each sequence, leading to a flurry of novel sequencing applications in functional genomics. Common challenges are associated with **shorter reads** and bioinformatics issues with the handling and analysis of massive amounts of data.

The three new technologies are outlined in Figure 1, 2 and 3, respectively, and comparatively analysed in Table 1, which highlights how their technological differences make them **complementary** in strengths and limitations, and consequently in applications. A reference to the latest Sanger platform stresses out their common advances ahead of the previously conventional sequencing technology.

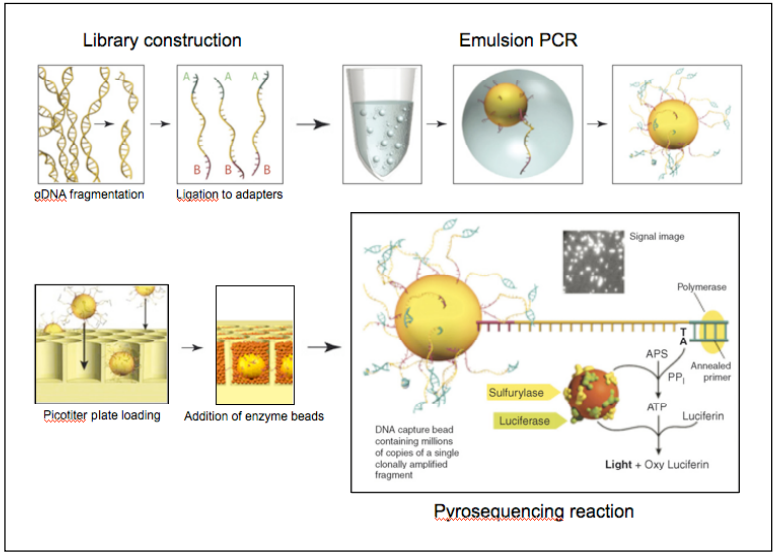

Figure 1. Genome Sequencer FLX System workflow. 1) Library construction: genomic DNA is fragmented, DNA fragments are ligated to adapters, denatured, and single-stranded fragments are immobilised on primer-coated beads by hybridisation, one single DNA molecule per bead. **2) Emulsion PCR:** Each bead is isolated with PCR reagents in distinct water droplets in an oil emulsion, resulting in the separate, clonal amplifications of unique DNA fragments on each bead. **3) Sequencing preparation:** The beads are loaded into a picotiter plate, one bead per well, and layered with enzyme beads. **4) Sequencing by synthesis with pyrosequencing readout:** the different types of nucleotides are flown sequentially over the picotiter plate; nucleotide incorporation by a DNA polymerase is detected by the emission of light induced by the release of pyrophosphate. After www.454.com/.

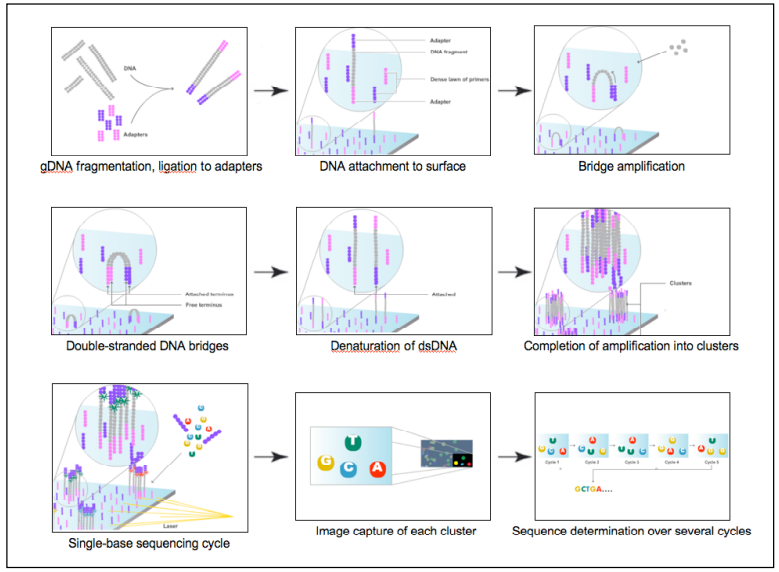

Figure 2. Illumina Genome Analyzer workflow. 1) Library construction: genomic DNA is fragmented, DNA fragments are ligated to adapters, denatured and randomly attached to a flow cell by hybridisation to pre-bound complementary primers**. 2) Solid-phase bridge PCR:** the immobilised DNA templates bend and hybridise to other primers bound to the flow cell in their vicinity and complementary to their free extremity. Amplification occurs in a bridge conformation; the strands straighten upon denaturation, and arch again for the next amplification, leading to local clusters of clonal DNA. **3) Sequencing by synthesis:** the cluster fragments are denatured, annealed to a sequencing primer and subjected to sequencing by synthesis using reversible base-specific fluorescent terminators (3' blocked nucleotides). At each round, unincorporated terminators are washed away, the sequence is determined by laser excitation, and the blocking 3' terminus and fluorophore are removed from the incorporated base for the next incorporation cycle. After www.illumina.com/.

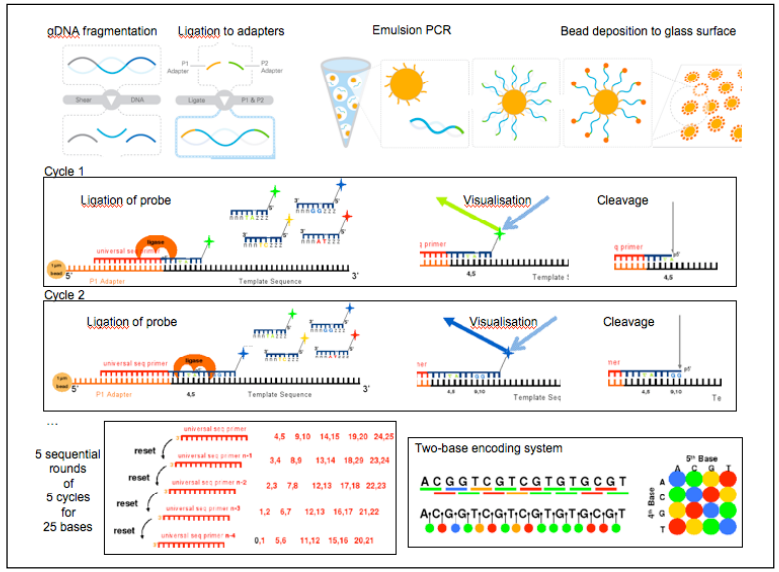

Figure 3. AB SOLiD workflow. 1) Library construction: genomic DNA is fragmented, fragments are ligated to adapters, **2) Emulsion PCR:** attachment of single DNA molecules on primer-coated beads, PCR in water-in-oil emulsion resulting in the separate, clonal amplifications of single DNA fragments on each bead. **3) Sequencing preparation:** the beads are randomly and densely deposited onto a glass slide. **4) Sequencing by ligation:** Hybridisation of a universal primer to the adaptor sequence, sequence-specific ligation of a fluorescent octamer probe, visualisation of fluorescence upon laser excitation, cleavage of the bases carrying the fluorescent group, new ligation in cycle 2, etc, the number of ligation cycles determining the length of the sequence read (here 5 cycles for 25 bases, 7 cycles for 35 bases in more recent versions). The ligation cycles effectively add 2 informative nucleotides at a time, at 5 bp interval. The template is reset after the last cycle, and a new round of cycles is initiated with the hybridisation of an n-1 universal primer. Five rounds (from n to n-4 universal primers) will fill in the sequence with overlapping dinucleotides, each base being read twice.

Each **octamer probe** consists of 3 degenerate bases (n), 1 dinucleotide determining the fluorescent dye, and 3 universal bases (z). The dinucleotide is positioned at the 4th and 5th bases of each octamer as depicted here, or at the 1st and 2^{d} bases in more recent versions. The fluorescent dye corresponding to the dinucleotide is attached to the 5' universal bases. There are 4 dyes and 16 dinucleotides in total, one dye representing 4 dinucleotides in a clever **two-base encoding system** that maximizes the discrimination between true polymorphism and sequencing error. The correct sequence out of the 4 possible reads inferred from the color code is deconvoluted thanks to the knowledge of the 5' base of the adaptor sequence, or base "0". After www3.appliedbiosystems.com.

Table 1. Comparative analysis of the currently competing sequencing technologies.
Specifications updated after October 2008 may be obtained at the appropriate websites.

	Latest Sanger platform	**Next-generation, massively parallel, sequencing technologies**		
	ABI 3730XL (Applied Biosystems) www.appliedbiosystems.com (released in 2003)	GS FLX System (Roche Diagnostics) www.454.com/ (released in October 2005)	Illumina Genome Analyzer (Illumina) www.illumina.com/ (released in June 2006)	SOLiD DNA Sequencer (Applied Biosystems) www3.appliedbiosystems.com (released in October 2007)
DNA template preparation	Fragmentation of genomic DNA or of large clones for clone-by-clone sequencing	Fragmentation of genomic DNA		
	DNA fragments cloned into bacterial vectors, transformed into *E. coli*. Bacterial selection, growth and plasmid purification. (Alternative to bacteria cloning: PCR-based method)	DNA fragments ligated to adapters and immobilised on solid support		
		Attachment of single-stranded DNA molecules on primer-coated sepharose **beads** (25-30 µm diameter), one DNA fragment per bead	Random attachment of single-stranded DNA fragments on primer-coated **silica matrix** (surface of flow cell)	Attachment of single-stranded DNA molecules on primer-coated magnetic **beads** (1 µm diameter), one DNA fragment per bead
DNA amplification		**Water-in-oil emulsion PCR** on beads, resulting in over 10 million copies of a unique DNA template per bead.	**Solid-phase bridge PCR** on the flow cell surface, resulting in a dense, random array of clonal clusters (up to 10 M clusters/cm^2, over 50 M clusters per channel, each cluster with 1000 template copies within 1 µm)	**Water-in-oil emulsion PCR** on beads, resulting in millions of copies of a unique DNA template per bead.
Sequencing platform	Capillary-based sequencer: electrophoresis-based separation of extension products in one capillary tube per template	**PicoTiter Plate™ Device**: fiber optic plate with 1.6 million 75-pL wells, filled with one DNA capture bead per well, layered with 1 µm-magnetic beads carrying the enzymes required for pyrosequencing. The plate is partitionable into 8 lanes.	**8-channel flow cell** carrying clonal DNA clusters from up to 8 different samples of randomly and densely distributed DNA clusters in each channel.	**Glass slide** onto which the beads are randomly and densely deposited. The slide consists of 2 independently controlled flow cells, each segmentable in up to 8 lanes.
Sequencing chemistry	**Sanger sequencing**: Sequencing by synthesis using **dideoxy chemistry** with **fluorescently labelled chain terminators** (ddNTPs). In separate thermocycling reactions, DNA strands complementary to the template are synthesised by a DNA polymerase. The synthesis stops each time a ddNTP is incorporated instead of a dNTP, resulting in a mixture of extension products of all sizes.	**Pyrosequencing**: (based on sequencing by synthesis) The four nucleotide types are flowed sequentially across the plate; the DNA polymerase-mediated incorporation of a nucleotide results in the release of pyrophosphate, converted to ATP, which fuels the luciferase-driven oxidation of luciferin and associated emission of light.	**Sequencing by synthesis**: the DNA polymerase sequentially incorporates **fluorescently labelled reversible terminators**, one at a time, specifically picked from a mixture of all 4 nucleotides. After laser excitation, the specific fluorescence is detected and recorded for each cluster over the flow cell. The 3' terminus blocking group and the fluorophore are then removed to allow the incorporation and detection of the next nucleotide.	**Sequencing by ligation** cycles of sequence-specific ligation, detection and cleavage of fluorescent octamer probes, which effectively add 2 nucleotide sequence at a time. Each nucleotide is read twice by overlapping dinucleotides from different octamers added in different rounds of ligation.
Sequencing readout	*After cleanup and resuspension, the products are loaded into the capillary sequencer.* Extension products are size-separated by electrophoresis through capillary tubes, up to a detection window where a laser excites the fluorescent labels of the chain terminators in the order of product size, and a detector produces an electropherogram of the nucleotide sequence.	**Real-time detection of nucleotide incorporation by chemiluminescence**, a CCD camera detecting the wells that emit light associated with luciferin oxidation. The nature of the incorporated nucleotide is deduced from the time at which the light is emitted, corresponding to a specific nucleotide flow through the plate. Because there is no termination system, several nucleotides are incorporated into one DNA strand in one flow at homopolymer	**Real-time detection of specific nucleotide fluorescent labels** from the reversible terminators by four-channel fluorescent scanning. The nucleotide sequence is directly deduced from the succession of the different fluorescent signals, each signal corresponding to a single nucleotide type at a single position.	**Real-time detection of fluorescent probes** of four colors, each corresponding to a set of 4 dinucleotides out of 16 in a 2-base encoding system. The correct succession of nucleotides out of the 4 possible sequences suggested by the color code is deconvoluted using knowledge of the adaptor sequence. (see SOLiD webinar on 2-base encoding system: marketing.appliedbiosystems.com/mk/get/SOLID_KNOWLEDGE_LANDING). Users are recommended to

		regions. The size of homopolymers will be deduced from the intensity of the signal at that moment. Errors occur when the signal becomes saturated (>3 bases).		perform all sequence analysis in "color space" before conversion to a sequence of nucleotides.
Read length	600-900 bp	400 bp	18, 26, 36 bp (50 bp available soon)	35 bp (25 bp in mate pairs) (50 bp planned for 2009)
Total length spanned by paired ends	Variable depending on insert sizes in cloning vectors	3 kb	200 bp (up to 600 bp)	3 kb
Nb reads per run	96	1.25 M	60 M (120 M for mate pairs)	120 M (240 M for mate pairs)
Data per run	**1 Mbp**	**500 Mbp**	**1.5 Gbp** **(3 Gbp for mate pairs)**	**4 Gbp** **(6 Gbp for mate pairs)**
Time per run	1 day	7 hours	2.5 days (5 days for mate pairs)	4 days (8 days for mate pairs)
Sequencing productivity	1 Mbp per day	1 Gbp per day	600 Mbp per day	1 Gbp per day
Machine cost	€400,000	€450,000	€500,000	€550,000
Run cost per base	~€1000/Mbp	~€20/Mbp ~50-fold cheaper than Sanger	~€5/Mbp ~200-fold cheaper than Sanger	
Bioinformatics capacity requirements	~ 1 MB of data	~ 15 GB of raw data	~1 TB in raw data, 300 MB in sequence results	~2 TB in raw data
Strengths	- Relatively long and accurate reads - efficient algorithms for assembly and annotation	- longest reads of the new sequencing technologies, - fastest run	Cheaper, higher throughput	- Similar price as Illumina per base, but over twice as much sequence and higher productivity despite longer runs. - Two-base encoding system increases base call reliability.
Limitations	- Expensive - low-throughput - labour intensive - cloning bias - no detection of rare mutations in heterogeneous sample	- Accuracy limitations: technical difficulty with long homopolymers resulting in indel errors; deteriorating sequence quality at 3' end - relatively short reads: issues with assembly of repetitive regions - still relatively expensive	- Accuracy limitations: Potential for substitution errors; deteriorating sequence quality at 3' end (after 20-25 bp); - Short reads: issues for assembly or mapping and annotation	- Accuracy limitations: negative bias for GC-rich regions - Short reads: issues for assembly or mapping and annotation - Adoption of color space (dinucleotide code) in sequence analysis
Applications of choice	Method of choice for *de novo* sequencing of complex genomes, and for routine fragment sequencing such as verifying clones or engineered constructs	Best Sanger competitor for ***de novo* sequencing projects** for small and simple genomes (bacterial, archeal, viruses, plastids, plasmids) and **metagenomics** projects. Best suitable length for **ancient DNA analysis** and 16S rRNA diversity analysis. **Resequencing** for genome-wide polymorphism discovery. In-depth analysis of whole transcriptomes in model and non-model organisms.	Best for **"seq-based" method** analyses applied to sequenced genomes, requiring higher number of reads rather than longer reads in transcriptomic and epigenomic analyses. Best for **resequencing** projects and **ultra deep SNP discovery** where lower costs for higher outputs outweigh the advantage of longer reads of GS FLX.	Should be in direct competition with Illumina applications. However, due to its recent launch, there are too few publications at this time to assess the range of SOLiD applications, and to compare its performance with competing technologies.
Publications	All genome sequencing projects prior to 2005 and still most current *de novo* sequencing projects.	www.454.com/news-events/publications.asp (~ 200 application papers, starting in 2005) - 45% in smaller genomes and metagenomic projects, - 45% in transcriptomic and epigenomic analysis - 10% in euk sequencing,	www.illumina.com/pages.ilmn?ID=87 (~50 application papers, starting in 2007) - Near 90% in transcriptome profiling and epigenomic analysis. - 8% in euk and 2% in smaller genome resequencing analysis.	marketing.appliedbiosystems.com/mk/get/SOLID_KNOWLEDGE_LANDING (3 application papers, starting May 2008) - Nucleosome position map, - transcriptome profiling, - bacterial genome resequencing.

3. Current and prospective applications of next-generation sequencing technologies:

In addition to the expected benefits of cost-effective DNA sequence information at revolutionary depth, scale and throughput, unexpected benefits have been derived from the new sequencing technologies: (1) the analysis of gene expression by transcriptomic profiling, and (2) the analysis of mechanisms behind the regulation of gene expression by epigenomic profiling. Altogether, novel applications in genome-wide genetic variation, transcriptomic and epigenomic analyses position the next-generation sequencing technologies as groundbreaking integrative tools providing unprecedented insights into **genome-wide functional genomics**.

Fundamental advances in genetics and genomics, transcriptomics and epigenomics have repercussions in virtually all fields of biology, with downstream applications in medicine and nutrition, plant and animal breeding and agriculture biotechnology. Furthermore, the new sequencing technologies have opened the gates to a new world of understanding of microbial diversity and ecology, with great promise for innovative applications in agriculture, nutrition, alternative energy production and the environment.

- GENOME LEVEL APPLICATIONS

1. *De novo* sequencing with the next-generation technologies

De novo genome sequencing refers to the sequencing of genomes for which there is no prior sequence information. To decipher a genome without prior information, individual reads must be assembled based on sequence overlaps only.

For small and simple genomes, the new sequencing technologies provide a fast and economic alternative to Sanger sequencing: the GS FLX System in particular, which produces the longest reads of the new technologies, has successfully generated *de novo* sequences of whole bacterial and archeal genomes.

For larger and more complex genomes, however, *de novo* sequence assembly represents a major bottleneck for the new technologies. Though laborious and expensive, Sanger technology remains better equipped for *de novo* sequencing of complex genomes, owing to the production of long reads and the possibility for clone-by-clone or hierarchical sequencing strategy. Nevertheless, the new sequencing technologies do bring a major breakthrough for unsequenced genomes of any size and complexity, with the possibility to generate cost-effective genome-wide sequence information for marker or gene discovery.

Applications:

DNA sequence is the foundation for genomics research, at the basis for rational breeding at the molecular level for improved yield, quality and sustainability of agricultural products. Furthermore, obtaining a whole genome sequence is a pre-

requisite to access a treasure trove of functional data generated by the whole-genome sequence-based applications of the new sequencing technologies.

1) Direct applications to whole-genome sequencing of simple genomes:
   - Sequencing of "exotic" organisms that may use novel metabolic pathways of interest, e.g., for adaptation to poor soils (Alcaraz *et al.*, 2008) or for converting biomass into energy (Hongoh *et al.*, 2008).
   - Sequencing of key uncultured bacteria/archae to provide reference genomes for genetic diversity analyses and to facilitate metagenomic analyses (Hongoh *et al.*, 2008).

2) Current applications to eukaryote genomes through the production of partially assembled yet informative sequences:
   - *De novo* transcriptome sequencing provides a first survey of exon sequences from uncharacterised genomes, instrumental for gene discovery, annotation, and SNP (single nucleotide polymorphism) discovery (Novaes *et al.*, 2008).
   - Direct sequencing of genomic fragments enable SNP discovery irrespective of gene expression in non-model organisms (Bekal *et al.*, 2008)

3) Foreseen applications, as eukaryote genomes become amenable to direct *de novo* sequencing by improved technologies:
   - Pursue the sequencing of model organisms as the basis for basic research in structural and functional genomics more cost-effectively.
   - Sequencing of major farm animals and crop plants to harness all the potential of molecular breeding
   - Sequencing of key evolutionary nodes in the phylogenetic tree for research in plant and animal evolution and domestication

2. Facilitated sequencing of related genomes and resequencing

Whole-genome sequencing is simplified when the genome of a closely related species is known. The related genome sequence is used as a reference genome or scaffold onto which short sequences can be aligned, greatly facilitating the new genome assembly. When the reference genome is from the same species as the genome to be sequenced, sequencing becomes a **resequencing** exercise, where the key to the variant genome reconstruction strictly lies in the correct mapping of the new reads to the reference genome.

Applications:

1) Exploitation of related genetic resources
   Facilitated sequencing of related species enables inter-specific comparative genomics for evolutionary, phylogenetic and functional analysis, and helps determine genetic targets for breeding at the molecular level, e.g.:
   - *vitis vinifera* grapevine genome, sequenced with Sanger technology, is used as a reference scaffold for high-throughput sequencing of the related species *vitis riparia*, of interest as a source of resistance to diseases (e.g., phylloxera).
   - Similarly, the genome sequence of cultivated tomato *Solanum lycopersicum* (previously *Lycopersicon esculentum*) could enable cost-effective sequencing of wild relative *S. pennellii (*previously *L. pennellii*), which would help reveal

the molecular nature of complex agronomic traits exposed in the introgression lines of *S. pennellii* chromosomal segments in *S. lycopersicum*.

2) Genetic diversity, ancient DNA and evolutionary studies:
   - resequencing multiple strains of model organisms to assess genetic diversity (Illumina proof-of-concept paper) (Hillier *et al.*, 2008)
   - Applications of ancient DNA analysis: high-throughput, cost-effective technologies allowing deep sequencing is essential for ancient DNA analysis considering it is often degraded and in minute concentration in specimen samples. Applications include:
     - resequencing ancient mitochondrial genomes, to reveal historical population dynamics (Gilbert *et al.*, 2008)
     - improved phylogenetic understanding, relationship between extinct and modern species
     - analysis of domestication process
     - impact analysis of past climatic changes on soil community and plant and animal distribution
     - the potential for "reactivation" of ancient genes by transgenesis

3) Genome-wide discovery of genetic variation as a pre-requisite for genotyping applications (e.g., population structure analysis, genotype-phenotype association studies, marker-assisted breeding, personal medicine, etc., developed in the dedicated Genotyping worksheet):
   - Foreseen application: resequencing entire core collections of plant and animal genetic resources to uncover most germplasm genetic diversity for subsequent genotyping and breeding applications.
   - Current applications to livestock and crops involve complexity-reduction strategies to reduce the size and cost of experiments:
     - deep resequencing of bovine reduced-representation libraries identifies large numbers of genome-wide SNPs in target populations (Van Tassell *et al.*, 2008)
     - resequencing expressed genes (ESTs) for cost-effective discovery of SNPs between two lines representing major heterotic groups of maize (Barbazuk *et al.*, 2007)
   - Application to catalog human genetic variation is ongoing with the 1000 genomes project (www.1000genomes.org).

4) Discovery of genetic variants associated with a phenotype, e.g.:
   - resequencing strains of *Mycobacterium tuberculosis* sheds light on antibiotic resistance (Andries *et al.*, 2005)
   - strain-to-reference comparison identifies markers for reactive biosecurity applications (La Scola *et al.*, 2008)
   - Whole-genome mutational profiling after mutagenesis breeding, e.g., on *Pichia stipitis* bred for improved xylose-to-ethanol conversion (Smith *et al.*, 2008)

5) Detection of rare somatic mutations by ultra-deep resequencing
   Successfully applied to the discovery of somatic mutations during tumor development (Campbell *et al.*, 2008), applications are foreseen to assess somaclonal mutations during clonal propagation of plants and trees.

3. Environmental sequence analysis (see Metagenomics worksheet)

Important breakthrough in microbiology and ecology, the new sequencing technologies have the capacity to analyse the DNA content of environmental samples, sequencing all microbes in a bulk, directly from their natural ecological context, irrespective of our ability to culture them. This novel application of sequencing with major outcomes for agriculture is treated in the dedicated Metagenomics worksheet.

- TRANSCRIPTOME LEVEL APPLICATIONS

Next-generation sequencing technologies can do more than providing raw sequence information: applied at the transcriptome level, they can replace microarray-based strategies and provide an open, digital platform for genome-wide analysis of gene expression, without relying on previous annotation data. Applied to a non-sequenced genome, transcriptome sequencing with long reads provides a first access to gene diversity. Applied to a sequenced genome, higher throughput short read sequencing provides deep quantitative analysis of gene expression on a genome scale.

1. *De novo* transcriptome sequencing for broad gene and marker discovery

Sequencing normalised full-length complementary DNA generated from pools of RNA extracted from diverse tissues, conditions and genotypes can provide a first survey of genetic diversity and a large set of genetic markers for species with no prior sequence information. This approach requires long reads to facilitate contig assembly and reconstitute transcripts sequence. Although generating shorter reads than Sanger, GS-FLX was shown to uncover more genetic diversity than ABI 3730.

Applications:
1) Transcript profiling without prior sequence knowledge
2) Gene sequence analysis of plants or animals whose genome complexity (large size, numerous repeated elements, polyploidy) excludes a genome project, as is the case for pea.
3) Production of new expressed sequences for gene discovery, functional analysis, annotation and SNP discovery (Novaes *et al.*, 2008).

2. Quantitative gene expression profiling

The so-called RNA-seq procedure enables deep quantitative analysis of gene expression of any sequenced genomes (Graveley, 2008). Here, the sequence is not the primary interest. In fact, knowing the sequence of the genome to be analysed is a prerequisite for RNA-seq. Sequence reads should be just long enough to be mapped on a reference genome. The interest is where the sequence maps, and how many reads map there, which provides hypothesis-free information on transcriptional units, splicing and expression levels. This approach is made possible due to the new technologies capacity to map and count reads following sequencing at great depth.

Applications:
1) Digitally measure the presence and prevalence of transcripts from known and previously unknown genes

2) Discovery of novel transcriptional units and alternative splicing
3) Possibility to distinguish sense from anti-sense transcripts
4) Precise measure of gene expression level and distinction between members of gene families
5) Deep sequencing to identify low-abundance transcripts
6) Cost-effective technique to profile transcriptomes in different mutants, different tissues, under different conditions (Lister et al., 2008; Wilhelm et al., 2008)

3. Deep sequencing of small RNAs

Specific to eukaryotes, small non-coding RNAs are key regulators of a number of biological processes including development, stress responses and genome stability. Small RNAs also play an important role in transgene expression.

There are different types of small RNAs, mostly belonging to two major groups: (1) the microRNAs (miRNAs), generated from broadly conserved *MIR* genes, which act as negative regulators of target genes by degrading, or inhibiting the translation of, complementary target mRNAs, and (2) the short or small interfering RNAs (siRNAs), generated from DNA repeats, transposons or incorrectly processed RNA transcripts, which are involved in gene silencing by guiding novel epigenetic modifications, or through mRNA cleavage.

Small RNAs can be specifically analysed in a dedicated smRNA-seq approach, applied to a fraction of total RNA, usually within a 15-30 nt interval. Specific bioinformatics pipelines enable to distinguish between types of small RNAs.

Applications:
1) Discovery of novel, mostly low-abundance, less-conserved miRNAs
   - organ-specific (e.g., rice grain, tomato fruit) (Moxon *et al.*, 2008; Zhu *et al.*, 2008)
   - lineage and/or species-specific (e.g., avian/chicken) (Glazov *et al.*, 2008)
   - abiotic or biotic stress-specific
2) Investigation of the role of small RNAs in development patterning, in lineage- and/or species-specific pathways and functions, in abiotic and biotic stress responses
3) Subsequent exploitation of miRNAs and RNA interference to modulate the expression of target genes of interest

- EPIGENOME LEVEL APPLICATIONS

Next-generation sequencing reveals genome-wide profiles of gene expression, but can it help understand how gene expression is regulated? The regulation of gene expression is partly controlled at the epigenetic level by modifications that affect gene expression without affecting the DNA sequence. Epigenetic modifications typically include DNA methylation, post-translational modification of histone proteins, and variations in nucleosome positioning. Unexpectedly, the next-generation sequencing technologies can not only detect but precisely map and quantify these modifications. Sequencing then enters new realms of investigation, that of (1) the fundamental mechanisms of regulation of gene expression by epigenetic modifications, (2) cell

differentiation and specialisation during normal development, and (3) the specific modulation of gene expression upon environmental triggers. These issues can now be addressed on a genome scale, across tissues, across treatments, developmental stages and generations, for a better understanding of how these modifications are acquired, orchestrated and inherited, for what consequences, and whether the ultimate code for these non-genetic modifications is not controlled in the DNA sequence after all.

1. DNA Methylation profiling

Cytosine methylation in specific sequence contexts causes stable and heritable gene silencing. DNA methylation can be analysed at single-base resolution by sequencing bisulphite-treated DNA, a procedure termed MethylC-seq or BS-seq, depending on the publications. Bisulphite treatment converts unmethylated cytosines to uracils, while methylated cytosines remain unchanged, enabling to deduce the status of each cytosine by straight sequencing. Other techniques assessing cytosine methylation exist that are currently cheaper but less exhaustive (using methylation-sensitive endonucleases), or less precise (using methyl-C immunoprecipitation).

Applications :
1) Reveal new methylation sites at single-base resolution, inform on global methylation patterning or specific methylated promoters (Cokus *et al.*, 2008)
2) Explore different pathways and regulation of methylation in different sequence context by profiling DNA methylation in different DNA methylation mutants (Cokus *et al.*, 2008)
3) Integrate genome-wide DNA methylation, small RNA, and mRNA profiles to analyse the global interplay of epigenetic modifications, RNA interference and transcription (Lister *et al.*, 2008).

2. Protein-DNA interaction profiling

The so-called ChIP-seq procedure enables to identify and quantify *in vivo* protein-DNA interactions on a genome scale. ChIP-seq combines high-throughput sequencing with chromatin immunoprecipitation (ChIP): sequencing DNA fragments immunoprecipitated with a DNA-binding protein of interest enables high-resolution mapping of binding sites to any sequenced genome (Figure 4).

DNA-binding proteins of interest include transcription factors and histone proteins. Histones may undergo a large range of post-translational modifications, which are proposed to function combinatorially or sequentially to regulate downstream functions, such as marking enhancer elements or guiding DNA methylation.

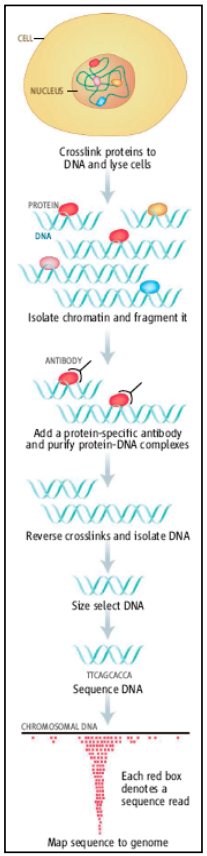

Figure 4. ChIP-seq workflow in Johnson, 2007 (from Fields, 2007)

Applications:

1) Genome-wide identification of DNA targets of transcription factors, for different cell types and physiological conditions (Johnson *et al.*, 2007)
2) Genome-wide profiling of binding sites of modified histones in different tissues, different conditions, or during differentiation of pluripotent stem cells to understand the epigenetic control of cell specialisation (Mikkelsen *et al.*, 2007)

3. Nucleosome positioning

The position of nucleosomes directly influences gene regulation by controlling access of transcription factors and transcription machinery to the DNA sequence. Various strategies involving sequencing mononucleosomal DNA with all three next-generation technologies enable to map the positions of nucleosomes at high resolution throughout the genome, giving unprecedented data sets for inferring positioning rules in relation to DNA sequence (Albert *et al.*, 2007).

Application: to acquire a better understanding of how genes are regulated by nucleosome positioning, and how nucleosome positioning is controlled.

4. Current limitations and challenges to overcome

1) The cost of the new sequencing technologies is still limiting

Proposed solutions:

1- Low-cost alternatives to whole-genome resequencing:
   To reduce the complexity and cost associated with sequencing an entire genome, a number of strategies enable to focus on specific regions of interest:
   - Transcriptome sequencing
   - Sequence capture technologies for sequencing a target region (candidate gene regions, BAC clone, exons, promoters, etc):
     - PCR amplification is currently the prevailing method for complexity reduction, but it is very expensive.
     - Microarray-based genomic selection is more flexible and cost-effective (Olson, 2007), with the caveat, however, that all hybridisation methods are unable to handle repeated sequences

2- Possibility to sequence multiple samples in on run:
   Multiplexing or barcoding samples with specific adaptors enable to distinguish an infinite number of samples in one run, over the limited number of sections inbuilt in the different technologies.

2) Challenges associated with short reads:

1- Genome assembly is impossible in repetitive regions when the repeats are longer than the length of the sequence reads.
2- Accurate mapping of short reads that align equally well with multiple locations (repetitive regions or dispersed gene families) is a significant challenge for the completion of resequencing and seq-based projects (RNA-seq, ChIP-seq).
3- Large genomic rearrangements are difficult to detect and characterise in resequencing projects.

Proposed solutions:

1- Deeper sequencing coverage than with Sanger is recommended to increase redundancy and compensate for short reads
2- New computer algorithms are generated to facilitate short read assembly (Butler *et al.*, 2008)
3- Paired-end mapping facilitates *de novo* assembly and genomic structural variation studies. Paired-end mapping technology is still under fine-tuning but has already proven successful (Korbel et al., 2007)
4- Hybrid strategies: the GS-FLX system is often used in combination with Sanger for *de novo* sequencing projects (Hongoh *et al.*, 2008), in proportions that should increase in favour of the new sequencing technology as the read lengths and sequencing accuracy increase with future technical and analytical software improvements.
5- Longer reads to come: some improvement in read length is expected in the current technologies and guaranteed in the coming 3rd generation technologies.

3) Sequencing accuracy: each technology has specific technological weaknesses

Distinguishing sequencing errors from true polymorphism is essential in resequencing projects, where genetic polymorphisms are expected between novel and reference sequences and specifically analysed for further genotyping applications.

1- The GS-FLX pyrosequencing technology has difficulty reading through homopolymers, resulting in frequent indel errors.
2- Illumina shows a risk for substitution errors and deteriorating sequence quality at the 3' end
3- SOLiD 2-base encoding system outperforms the other technologies and nears perfection in SNP discovery due to error elimination based on dinucleotide code (Smith *et al.*, 2008).

Proposed solution: combine different sequencing technologies to compensate for their different types of weaknesses.

4) Bioinformatics issues:

1- Long term challenges with the storage and management of Terabytes of data
2- Need for new software and algorithms for sequence assembly, annotation, analysis, and comparison. As an example, specific issues are associated with the so-called "multireads", sequence reads that map equally well to several places in a genome, as opposed to "unireads", which map unequivocally to one locus. Several algorithms are currently developed to take into account information coming from multi-reads rather than focusing only on unireads.
3- Issues with adoption of SOLiD: users are recommended to make all sequence analysis in color space (SOLiD dinucleotide code). Published nucleic databases must be converted for color space analysis.

5) Additional challenges for transcriptomic profiling:

1- Issues with classification of unexpected reads, falling outside annotated transcript boundaries (DNA contamination or true RNA? New transcriptional unit or belonging to adjacent unit? etc)
2- Issues with sampling heterogeneous transcriptomes across or within a tissue

6) **Additional challenges for epigenomic profiling:**
   1- ChIP experiments rely on the quality and specificity of antibodies
   2- Because epigenetic modifications may either be inherited from a previous generation, programmed during cell differentiation, or induced upon environmental triggers, every organism is a mosaic of different epigenomes. As a result, an epigenomic profile is likely to represent a mix of different chromatin status from different cells.
   3- Cross-linking methods must be optimised for access to specific cell types

Next-next-generation in the pipeline: a number of companies are developing third- or next-next-generation sequencing technologies based on single-molecule analysis (no amplification step), using different strategies such as high-definition optics, FRET-based approach, zero-mode waveguides or hybridisation-assisted nanopore sequencing, promising higher throughput, lower costs, and longer reads (over 1 kb). Some of these technologies will be launched in 2009, making the strategic choice for an investment in sequencing technologies all the more difficult for sequencing centres and individual laboratories.

5. Glossary

DNA: deoxyribonucleic acid, the molecule carrying genetic information.

Base, base pair: each of the four subunits, or nucleotides, composing DNA: adenine (A), guanine (G), cytosine (C), and thymine (T). In the DNA molecule, complementary bases (A/T and G/C) are linked in pairs across two chains of nucleotides forming a double-stranded molecule. The terms "base" and "base pair" are often used interchangeably because the knowledge of the base of one strand is sufficient to infer the corresponding base on the other strand.

DNA sequencing: determining the order or sequence of bases in DNA.

Genomics: field of study of the genome, i.e. the complete genetic make-up of an organism.

Comparative genomics: comparative analysis of the sequences of two or more related organisms, providing insight into genome evolution and gene function.

Phylogenetics: field of study of evolutionary relatedness among various groups of organisms.

Functional genomics: analysis of the function of genes, by getting information on biological activity, but also on gene polymorphism, gene expression and regulation.

Metagenomics, whole-community genomics or environmental genomics: field of study of the metagenome, defined as all the genomes of a microbe community in a particular environment.

Transcriptomics: simultaneous analysis of many transcripts, aiming at the full complement of transcripts (the transcriptome) of an individual, an organ, a tissue or a cell.

Epigenetic modifications: Developmentally programmed or environmentally induced chemical "marking" of chromatin (DNA and associated histone proteins), which affects gene expression without affecting the DNA sequence. Epigenetic modifications are heritable from cell to cell through mitosis, or from generation to generation through meiosis. Modifications typically include DNA methylation at some cytosine residues (CpG dinucleotides in animals), post-translational methylation, acetylation, phosphorylation or ubiquitylation of histone proteins, and change in nucleosome density and positioning.

Epigenomics: simultaneous analysis of many epigenetic modifications, aiming at the full complement of epigenetic modifications (epigenome) in an individual, an organ, a tissue, or a cell.

Pyrosequencing: Type of sequencing by synthesis where the addition of a nucleotide is detected by an emission of light, driven by a luciferin/luciferase system.

Homopolymer: an uninterrupted stretch of a single nucleotide.

Indel: genetic variant consisting of an insertion or deletion of a sequence in a genetic locus of one individual compared to another. The words are fused into one because an insertion can be viewed as a deletion in the reciprocal comparison.

Paired-end reads, mate pairs: sequences from both ends of a DNA molecule, separated by a known distance, which allow to compensate for the limitations due to short read lengths in sequence assembly.

***de novo* sequencing:** sequencing without prior sequence information. Sequence assembly is based on overlaps in sequence reads.

Resequencing: sequencing using a reference sequence from the same species as a scaffold onto which reads can be aligned instead of having to assemble the novel sequence based on sequence overlaps. Resequencing enables to uncover and profile genetic variation in a given species.

Whole-genome shotgun strategy: the whole genome is fragmented in a bulk, fragments are randomly sequenced and subsequently assembled back in chromosomes based on overlapping sequences. This approach does not require prior physical map information but requires heavy computational work.

Clone-by-clone or hierarchical sequencing: the genome is first fragmented in large clones, which are physically ordered along a genetic map and a minimal tiling path of overlapping clones are selected for separate sequencing.

Single-Nucleotide Polymorphism (SNP): most common type of genetic variant, consisting of a single nucleotide difference between two individuals at a particular site in the DNA sequence. SNPs can be used as genetic markers.

Ancient DNA: DNA recovered from any *post mortem* material, including preserved biological remains, ice and permafrost material. Ancient DNA is characterised by low concentration in biological material, degradation into short fragments, and typical chemical modifications from oxidative damage and hydrolytic processes.

RNAs: ribonucleic acids, small molecules resulting from the transcription of DNA units. A portion of RNAs code for proteins, chains of amino-acids translated from nucleotides. Protein-coding RNAs are called **messenger RNAs** or **mRNAs**. Non-coding RNAs are mostly involved in the translation machinery of mRNAs into proteins and in the regulation of gene expression.

RNA splicing, alternative splicing: mechanism of maturation of eukaryotic mRNA after transcription, consisting of the removal of some sequences (the introns) and the fusion of the remaining sequences (the exons) resulting into mature mRNAs that will be translated into proteins. Although splicing is predetermined by specific sequence motifs, a single mRNA can yield a range of proteins depending on alternative splicing processes by which introns can be retained or exons skipped.

Small RNAs: some small RNAs, mostly between 21 and 24 nucleotide-long, are non-coding RNAs involved in regulating gene expression during development or in response to environmental factors. There are two major classes of small non-coding RNAs: the miRNAs and siRNAs.

MicroRNAs (miRNAs) (21-24 nt RNAs derived from single-stranded precursor RNAs fold back in hairpin structures from *bona fide MIR* genes): sequence-specific post-transcriptional negative regulators, which degrade, or inhibit the translation of, complementary target mRNAs.

Short or small interfering (siRNAs) (21-24 nt RNAs derived from double-stranded RNA generated from DNA repeats, including transposon and retroelement sequences, or incorrectly processed RNA transcripts): involved in *de novo* gene silencing at the transcriptional level through guiding epigenetic modifications (initiating a certain type of DNA methylation, a number of histone post-translational modifications and changes in nucleosome positioning) but also at the post-transcriptional level by guiding mRNA cleavage.

RNA-seq: quantitative gene expression profiling by high-throughput sequencing of complementary DNA resulting from the reverse transcription of RNA. The transcriptional units are defined by mapping the resulting sequence reads onto a genome sequence. The level of expression is digitally measured by counting the number of reads per transcriptional unit.

ChIP: Chromatin immunoprecipitation using antibodies specific to DNA-binding proteins.

ChIP-seq: Technique combining chromatin immunoprecipitation and high-throughput sequencing to identify and map specific protein-DNA interactions genome wide.

BS-seq, MethylC-seq: Sequencing following bisulphite treatment to profile cytosine methylation.

Chromatin: DNA and associated proteins, including histones.

Histones: Proteins around which DNA winds. Histones are essential for DNA compaction, and play a role in gene regulation through a large number of posttranslational modifications.

Nucleosome: fundamental structural unit of chromatin: 146-147 base pairs of DNA wrapped around a histone octamer.

Bisulphite treatment: used to map methylated cytosines, this treatment prevents changes to the methylation status of the sample, and induces the deamination of unmethylated cytosines to uracils, while methylated cytosines remain unchanged.

6. Key references

Albert, I., Mavrich, T. N., Tomsho, L. P., Qi, J., Zanton, S. J., Schuster, S. C., and Pugh, B. F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446, 572-576.

Alcaraz, L. D., Olmedo, G., Bonilla, G., Cerritos, R., Hernandez, G., Cruz, A., Ramirez, E., Putonti, C., Jimenez, B., Martinez, E., *et al.* (2008). The genome of Bacillus coahuilensis reveals adaptations essential for survival in the relic of an ancient marine environment. Proc Natl Acad Sci U S A 105, 5803-5808.

Andries, K., Verhasselt, P., Guillemont, J., Gohlmann, H. W., Neefs, J. M., Winkler, H., Van Gestel, J., Timmerman, P., Zhu, M., Lee, E., Williams, P., de Chaffoy, D., Huitric, E., Hoffner, S., Cambau, E., Truffot-Pernot, C., Lounis, N., and Jarlier, V. (2005). A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Sci 307, 223-227.

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., and Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. Plant J 51, 910-918.

Bekal, S., Craig, J. P., Hudson, M. E., Niblack, T. L., Domier, L. L., and Lambert, K. N. (2008). Genomic DNA sequence comparison between two inbred soybean cyst

nematode biotypes facilitated by massively parallel 454 micro-bead sequencing. Mol Genet Genomics 279, 535-543.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18, 810-820.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A., and Stratton, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A 105, 13081-13086.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215-219.

Gilbert, M. T., Kivisild, T., Gronnow, B., Andersen, P. K., Metspalu, E., Reidla, M., Tamm, E., Axelsson, E., Gotherstrom, A., Campos, P. F., *et al.* (2008). Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. Sci 320, 1787-1789.

Glazov, E. A., Cottee, P. A., Barris, W. C., Moore, R. J., Dalrymple, B. P., and Tizard, M. L. (2008). A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res 18, 957-964.

Graveley, B. R. (2008). Molecular biology: power sequencing. Nature 453, 1197-1198.

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., *et al.* (2008). Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5, 183-188.

Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Taylor, T. D., Kudo, T., Sakaki, Y., Toyoda, A., Hattori, M., and Ohkuma, M. (2008). Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. Proc Natl Acad Sci U S A 105, 5555-5560.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Sci 316, 1497-1502.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Sci 318, 420-426.

La Scola, B., Elkarkouri, K., Li, W., Wahab, T., Fournous, G., Rolain, J.-M., Biswas, S., Drancourt, M., Robert, C., Audic, S., Lofdahl, S., and Raoult, D. (2008). Rapid comparative genomic analysis for clinical microbiology: The Francisella tularensis paradigm. Genome Res 18, 742-750.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133, 523-536.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., *et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553-560.

Moxon, S., Jing, R., Szittya, G., Schwach, F., Rusholme Pilcher, R. L., Moulton, V., and Dalmay, T. (2008). Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. Genome Res., in press.

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Jr., Grattapaglia, D., Sederoff, R. R., and Kirst, M. (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9, 312.

Olson, M. (2007). Enrichment of super-sized resequencing targets from the human genome. Nat Methods 4, 891-892.

Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P., *et al.* (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res., in press.

Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., and Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5, 247-252.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453, 1239-1243.

Zhu, Q. H., Spriggs, A., Matthew, L., Fan, L., Kennedy, G., Gubler, F., and Helliwell, C. (2008). A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Res 18, 1456-1465.

7. Consulted expert

Patrick Wincker
Génoscope
Centre National de Séquençage
2, rue Gaston Crémieux
CP5706 91057 Evry Cedex