Institut de Recherche pour le Développement
FRANCE

NGS sequence application, a few examples

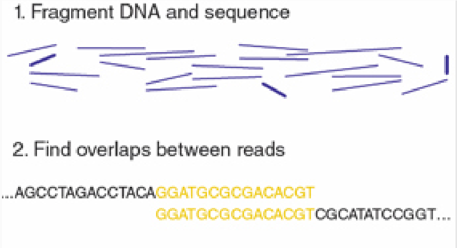
Dr Francois Sabot & Christine Tranchant-Dubreuil

8th of October, 2018

IRD - UMR DIADE

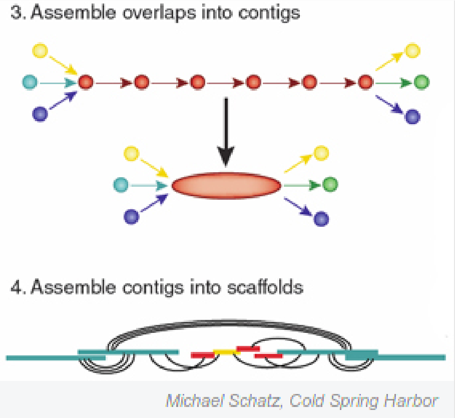
Analyses

Assembly


1. Fragment DNA and sequence
2. Find overlaps between reads
...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

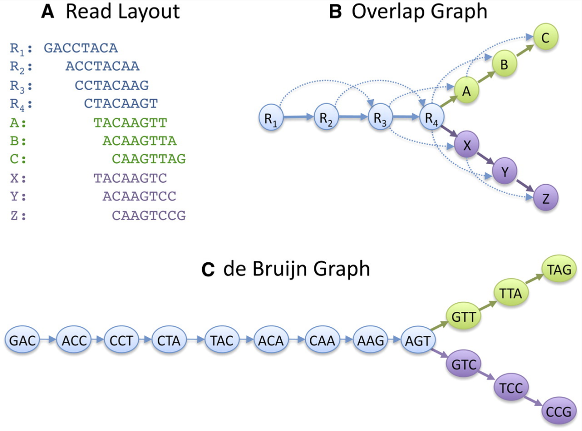
From Baker, 2012

Assembly


3. Assemble overlaps into contigs
4. Assemble contigs into scaffolds
Michael Schatz, Cold Spring Harbor

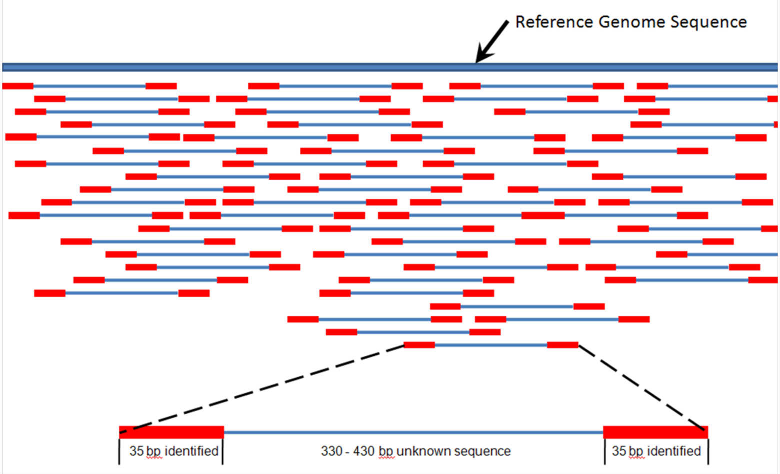
From Baker, 2012

Assembly



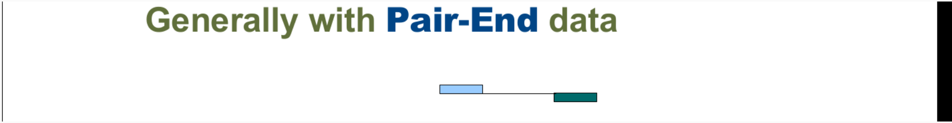
From Schatz, 2010

Mapping



From Wikipedia

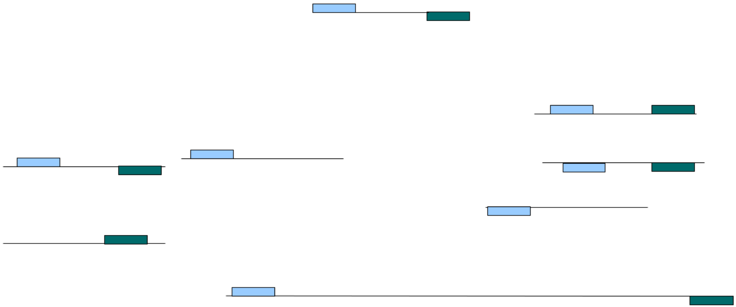
Mapping approaches for differences



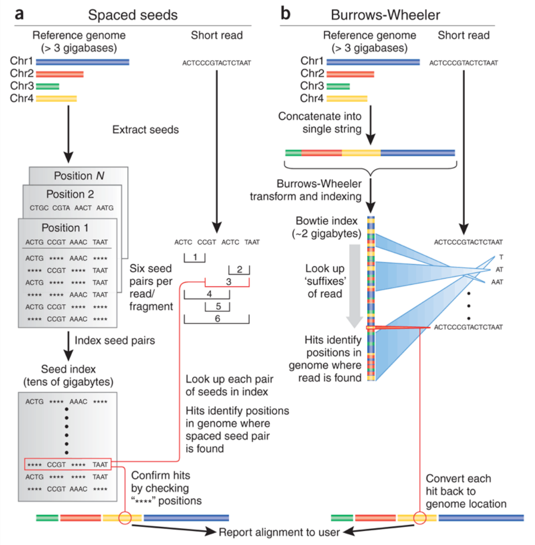
Generally with **Pair-End** data

Mapping approaches for differences

Generally with Pair-End data

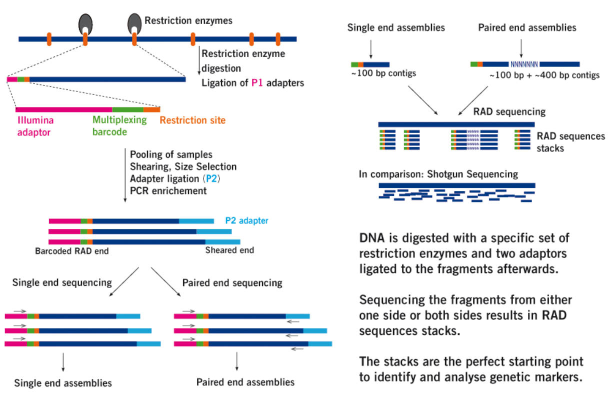


Mapping



From Trapnell & Salzberg, 2006

RAD-Seq



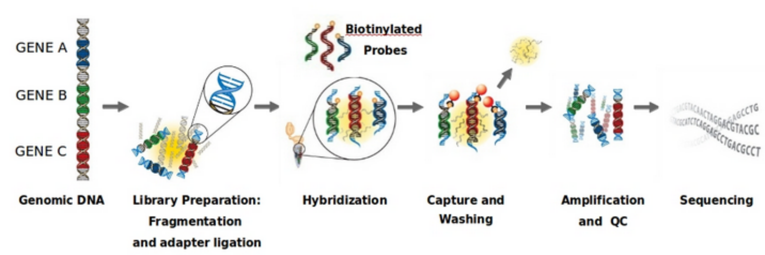
DNA is digested with a specific set of restriction enzymes and two adaptors ligated to the fragments afterwards.

Sequencing the fragments from either one side or both sides results in RAD sequences stacks.

The stacks are the perfect starting point to identify and analyse genetic markers.

From Eurofins

CaptureSeq



From CGFB, Bordeaux, France

Quantification

- Mainly in RNA sequencing, but also in CNV (Copy Number Variation)

Quantification

- Mainly in RNA sequencing, but also in CNV (Copy Number Variation)
- Counting the number of reads/bases at each position

Quantification

- Mainly in RNA sequencing, but also in CNV (Copy Number Variation)
- Counting the number of reads/bases at each position
- More precise than ChiP

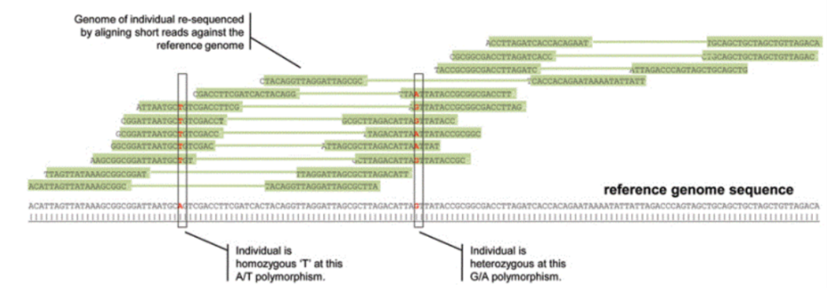
Quantification

- Mainly in RNA sequencing, but also in CNV (Copy Number Variation)
- Counting the number of reads/bases at each position
- More precise than ChiP
- Need to be reproduced

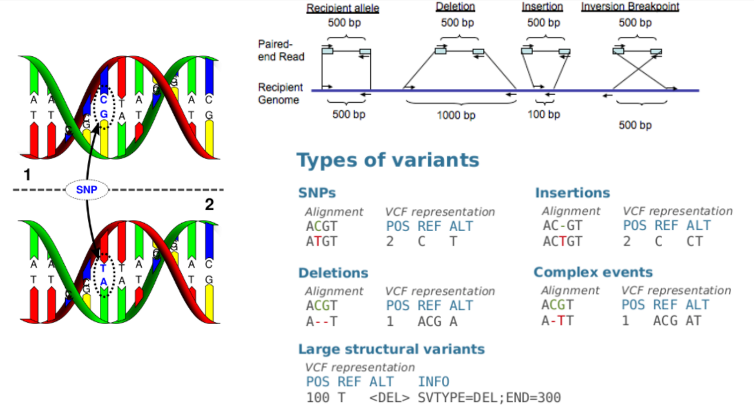
Quantification

- Mainly in RNA sequencing, but also in CNV (Copy Number Variation)
- Counting the number of reads/bases at each position
- More precise than ChiP
- Need to be reproduced
- Lots of Statistical models and Controls behind

SNP and InDel Detection


Genome of individual re-sequenced by aligning short reads against the reference genome
reference genome sequence
Individual is homozygous 'T' at this A/T polymorphism.
Individual is heterozygous at this G/A polymorphism.

SNP and InDel Detection



Types of variants

SNPs

```
Alignment   VCF representation
ACGT        POS REF ALT
ATGT        2   C   T
```

Insertions

```
Alignment   VCF representation
AC-GT       POS REF ALT
ACTGT       2   C   CT
```

Deletions

```
Alignment   VCF representation
ACGT        POS REF ALT
A--T        1   ACG A
```

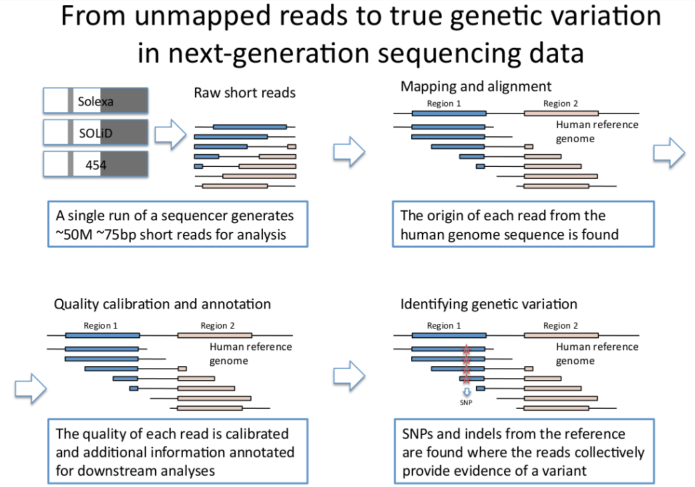
Complex events

```
Alignment   VCF representation
ACGT        POS REF ALT
A-TT        1   ACG AT
```

Large structural variants

```
VCF representation
POS REF ALT   INFO
100 T   <DEL> SVTYPE=DEL;END=300
```

From raw to refined data...



From unmapped reads to true genetic variation in next-generation sequencing data

Raw short reads

A single run of a sequencer generates ~50M ~75bp short reads for analysis

Mapping and alignment

The origin of each read from the human genome sequence is found

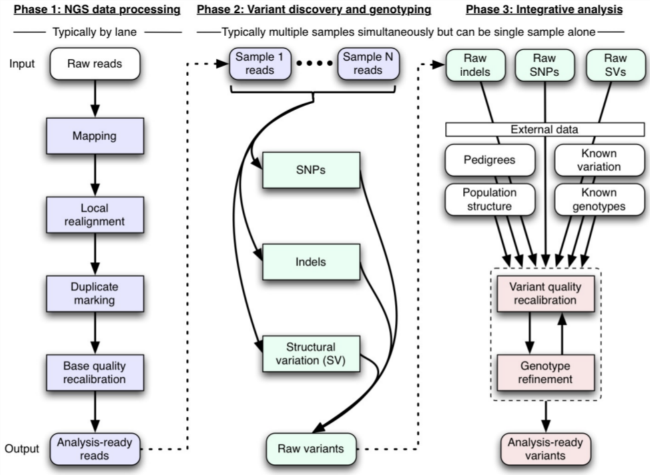
Quality calibration and annotation

The quality of each read is calibrated and additional information annotated for downstream analyses

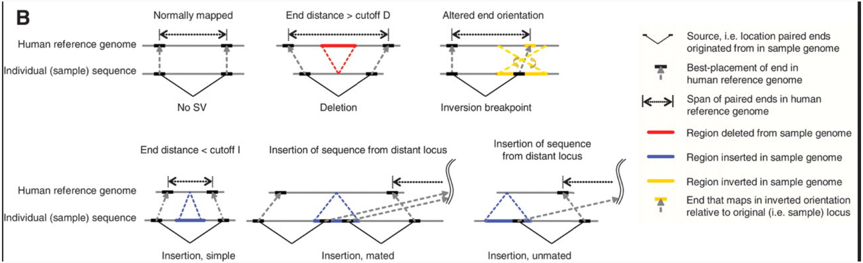
Identifying genetic variation

SNPs and indels from the reference are found where the reads collectively provide evidence of a variant

And even more refined...

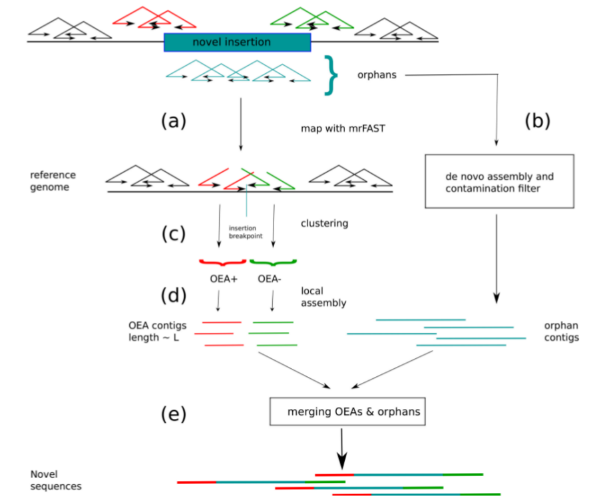


Structural Variant Detection

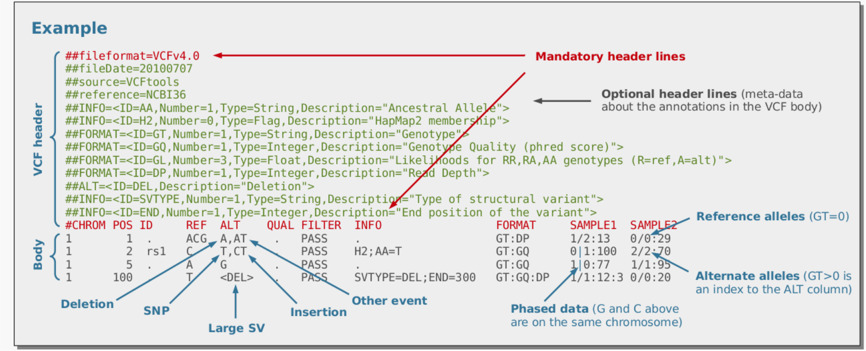

B
Normally mapped
End distance > cutoff D
Altered end orientation
Human reference genome
Individual (sample) sequence
No SV
Deletion
Inversion breakpoint
End distance < cutoff I
Insertion of sequence from distant locus
Insertion of sequence from distant locus
Human reference genome
Individual (sample) sequence
Insertion, simple
Insertion, mated
Insertion, unmated
Source, i.e. location paired ends originated from in sample genome
Best-placement of end in human reference genome
Span of paired ends in human reference genome
Region deleted from sample genome
Region inserted in sample genome
Region inverted in sample genome
End that maps in inverted orientation relative to original (i.e. sample) locus

From Korbel et al, 2007

Structural variation, an approach



Common File for all Variations, the VCF



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT     QUAL FILTER INFO                FORMAT     SAMPLE1    SAMPLE2
1        1 .      ACG  A,AT    .    PASS   .                   GT:DP      1/2:13     0/0:29
1        2 rs1    C    T,CT    .    PASS   H2;AA=T             GT:GQ      0|1:100    2/2:70
1        5 .      A    G       .    PASS   .                   GT:GQ      1|0:77     1/1:95
1      100 .      T    <DEL>   .    PASS   SVTYPE=DEL;END=300  GT:GQ:DP   1/1:12:3   0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

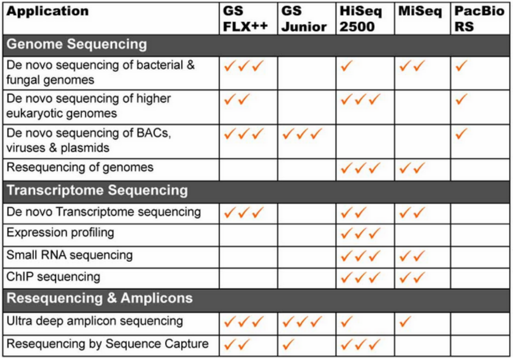
Large SV

Insertion

Other event

VCF = Variant Call Format From 1000 Genomes Project

Which Technology for Which application ?



Application	GS FLX++	GS Junior	HiSeq 2500	MiSeq	PacBio RS
Genome Sequencing					
De novo sequencing of bacterial & fungal genomes	✓✓✓		✓	✓✓	✓
De novo sequencing of higher eukaryotic genomes	✓✓		✓✓✓		✓
De novo sequencing of BACs, viruses & plasmids	✓✓✓	✓✓✓			✓
Resequencing of genomes			✓✓✓	✓✓	
Transcriptome Sequencing					
De novo Transcriptome sequencing	✓✓✓		✓✓	✓✓	
Expression profiling			✓✓✓		
Small RNA sequencing			✓✓✓	✓✓	
ChIP sequencing			✓✓✓	✓✓	
Resequencing & Amplicons					
Ultra deep amplicon sequencing	✓✓✓	✓✓✓	✓	✓	
Resequencing by Sequence Capture	✓✓	✓	✓✓✓		

From Eurofins

Which Technology for Which application ?

From my own personal Experience:

Assembly : Nanopore, PacBio, Illumina (MySeq + HiSeq, various libraries)

Which Technology for Which application ?

From my own personal Experience:

Assembly : Nanopore, PacBio, Illumina (MySeq + HiSeq, various libraries)

SNP detection : Illumina

Which Technology for Which application ?

From my own personal Experience:

Assembly : Nanopore, PacBio, Illumina (MySeq + HiSeq, various libraries)

SNP detection : Illumina

SV Variation : Nanopore, PacBio, Illumina, IonTorrent

Which Technology for Which application ?

From my own personal Experience:

Assembly : Nanopore, PacBio, Illumina (MySeq + HiSeq, various libraries)

SNP detection : Illumina

SV Variation : Nanopore, PacBio, Illumina, IonTorrent

Quantification : Illumina

Limits, Technical

- Amount of original samples

Limits, Technical

- Amount of original samples
- Size of sequenced unit

Limits, Technical

- Amount of original samples
- Size of sequenced unit
- Error rate

Limits, Technical

- Amount of original samples
- Size of sequenced unit
- Error rate
- Volume of Outputted data

Limits, Technical

- Amount of original samples
- Size of sequenced unit
- Error rate
- Volume of Outputted data

All linked to technical constraints

Limits, Technical

- Cleaning data level

Limits, Technical

- Cleaning data level
- Mapping Conditions

Limits, Technical

- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions

Limits, Technical

- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions
- Variation Calling level

Limits, Technical

- Cleaning data level
- Mapping Conditions
- Mapping Cleaning Conditions
- Variation Calling level

All linked to the Specificity/Sensitivity Informatics Paradox

Limits, Biological

- Availability of Sample

Limits, Biological

- Availability of Sample
- Choice of Sample

Limits, Biological

- Availability of Sample
- Choice of Sample
- Amount of Sample

Limits, Biological

- Availability of Sample
- Choice of Sample
- Amount of Sample
- Purity of Sample

Limits, Biological

- Availability of Sample
- Choice of Sample
- Amount of Sample
- Purity of Sample
- Size of sample (for Assembly/Mapping essentially)

Applications

Population Genomics/Genomics

- Gene discovery/GWAs

Population Genomics/Genomics

- Gene discovery/GWAs
- Species Definition

Population Genomics/Genomics

- Gene discovery/GWAs
- Species Definition
- Subspecies/specific subgroup definition

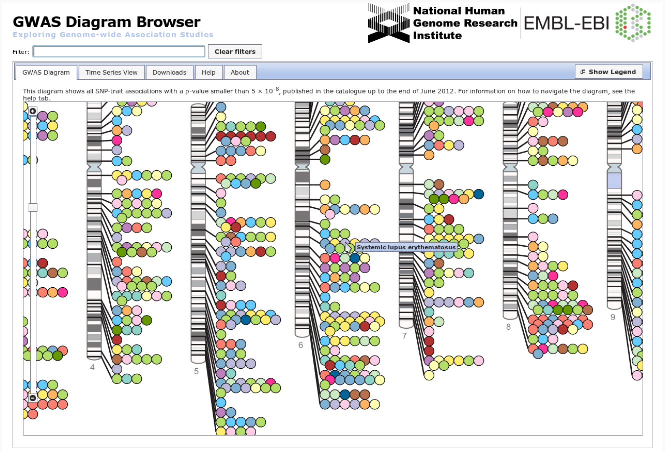
Population Genomics/Genomics

- Gene discovery/GWAs
- Species Definition
- Subspecies/specific subgroup definition
- Global genotyping (for breeding in agriculture e.g.)

Population Genomics/Genomics

- Gene discovery/GWAs
- Species Definition
- Subspecies/specific subgroup definition
- Global genotyping (for breeding in agriculture e.g.)
- Genomic Ecology (Transposable elements, etc...)

Example in GWAs & Population Genomics



GWAS Diagram Browser

Exploring Genome-wide Association Studies

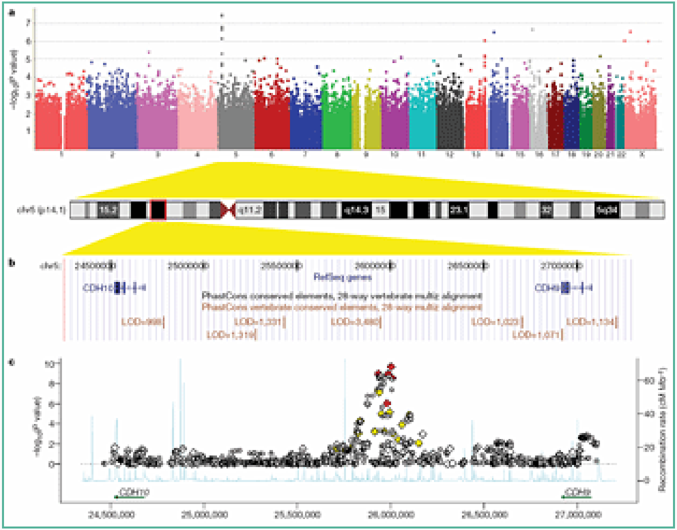
Filter: Clear filters

GWAS Diagram | Time Series View | Downloads | Help | About | Show Legend

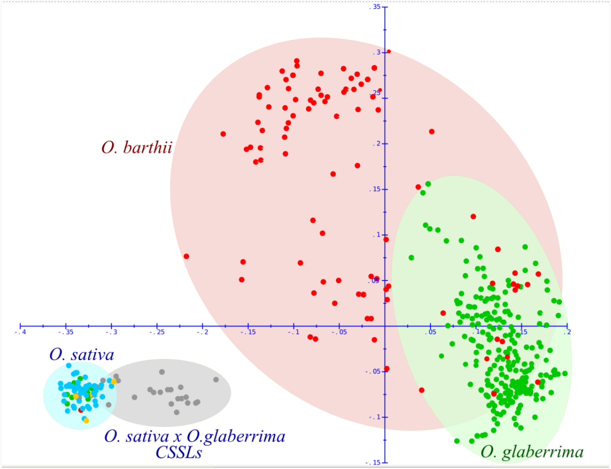
This diagram shows all SNP-trait associations with a p-value smaller than 5 × 10^{-8}, published in the catalogue up to the end of June 2012. For information on how to navigate the diagram, see the help tab.

National Human Genome Research Institute
EMBL-EBI
4
5
6
7
8
9
Systemic lupus erythematosus

Example in GWAs & Population Genomics

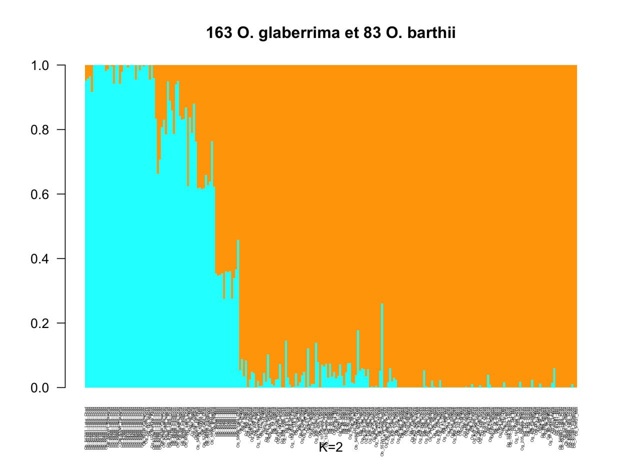


Example in Global Genotyping & Population Genomics



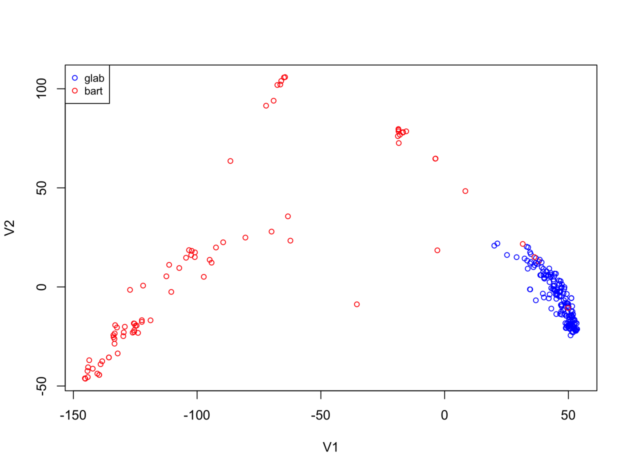
From Orjuela et al, 2014

Example in Global Genotyping & Population Genomics



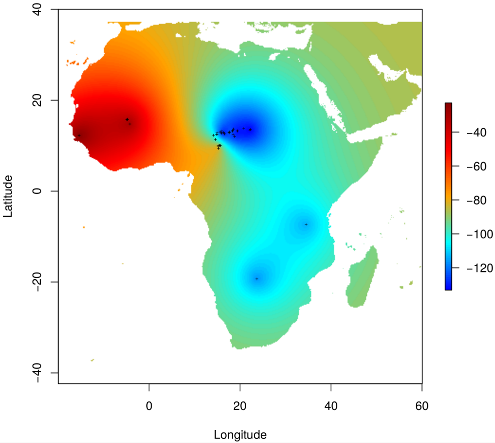
From Cubry et al, 2018

Example in Global Genotyping & Population Genomics



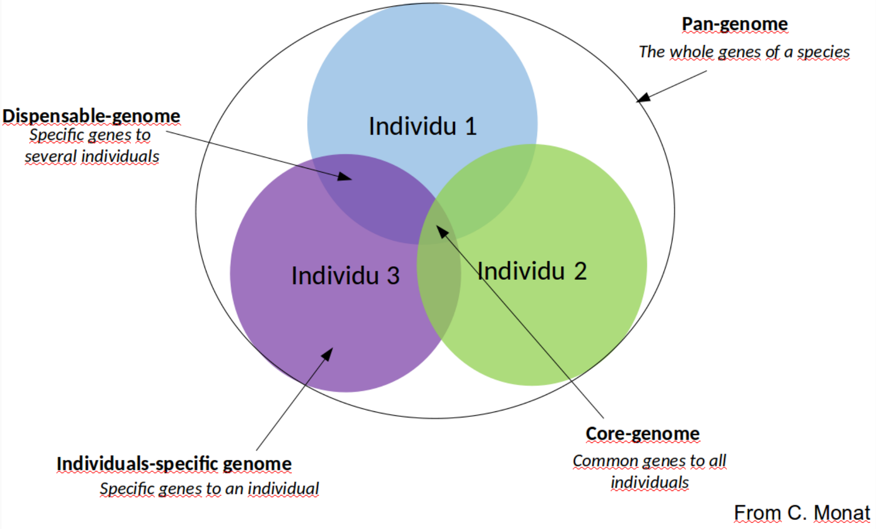
From Cubry et al, 2018

Example in Global Genotyping & Population Genomics



From Cubry et al, 2018

Pangenomic



Pan-genome

The whole genes of a species

Dispensable-genome

Specific genes to several individuals

Individuals-specific genome

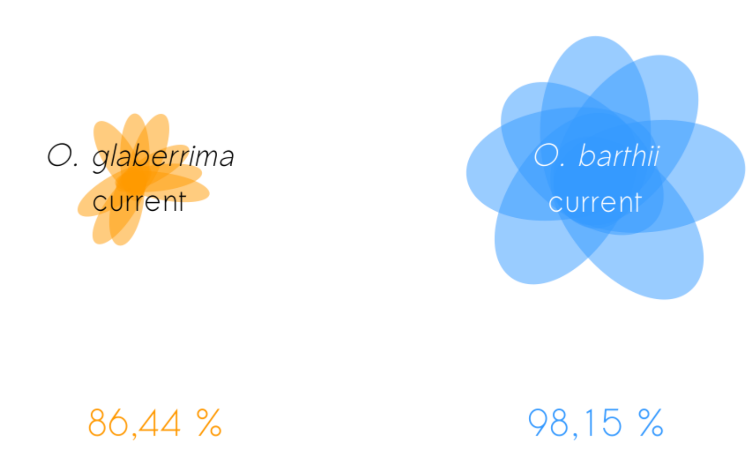
Specific genes to an individual

Core-genome

Common genes to all individuals

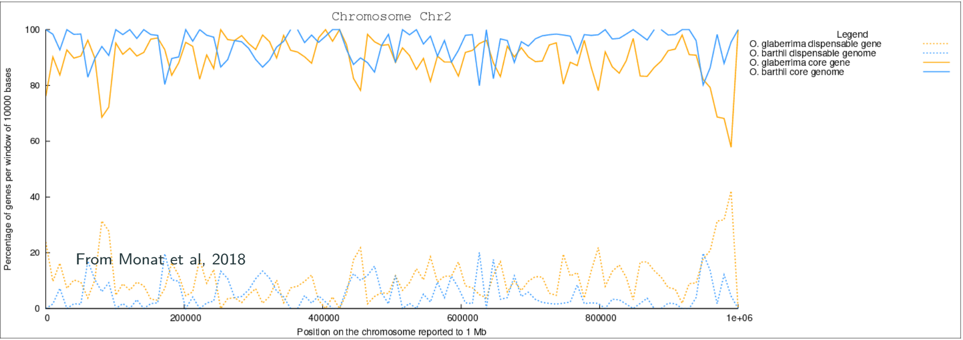
From C. Monat

Pangenomic



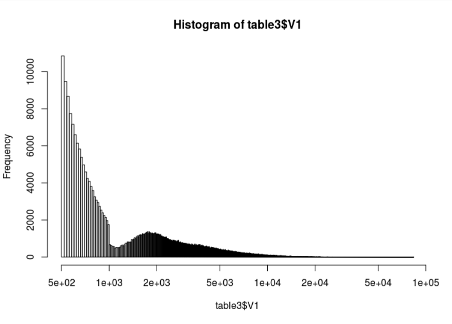
From Monat et al, 2016

Pangenomic



From Monat et al, 2018

Some really recent results...



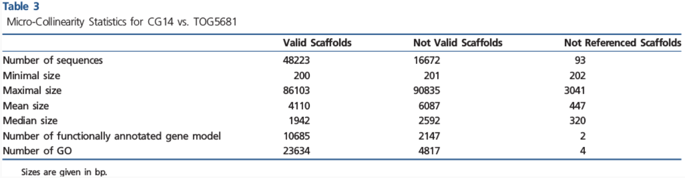


Table 3

Micro-Collinearity Statistics for CG14 vs. TOG5681

	Valid Scaffolds	Not Valid Scaffolds	Not Referenced Scaffolds
Number of sequences	48223	16672	93
Minimal size	200	201	202
Maximal size	86103	90835	3041
Mean size	4110	6087	447
Median size	1942	2592	320
Number of functionally annotated gene model	10685	2147	2
Number of GO	23634	4817	4

Sizes are given in bp.

From Monat et al, 2017

Transcriptomics

- Level of expression in different conditions or in different individuals

Transcriptomics

- Level of expression in different conditions or in different individuals
- Variation in sequences

Transcriptomics

- Level of expression in different conditions or in different individuals
- Variation in sequences
- Variation of splicing

Transcriptomics

- Level of expression in different conditions or in different individuals
- Variation in sequences
- Variation of splicing
- Variation of editing

Transcriptomics

- Level of expression in different conditions or in different individuals
- Variation in sequences
- Variation of splicing
- Variation of editing
- Detection of putative coding/active sequence

smallRNA Transcriptomics

- Level of expression in different conditions or in different individuals

smallRNA Transcriptomics

- Level of expression in different conditions or in different individuals
- Variation in sequences

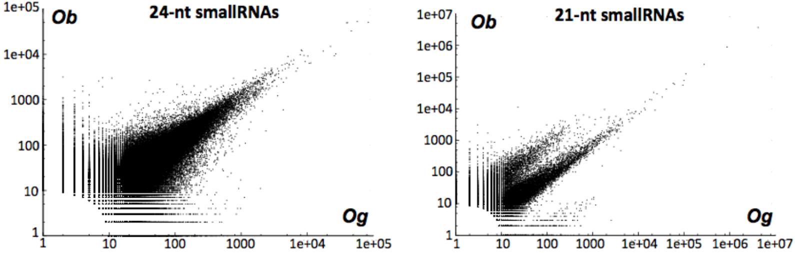
smallRNA Transcriptomics

- Level of expression in different conditions or in different individuals
- Variation in sequences
- Variation in specific forms

smallRNA Transcriptomics

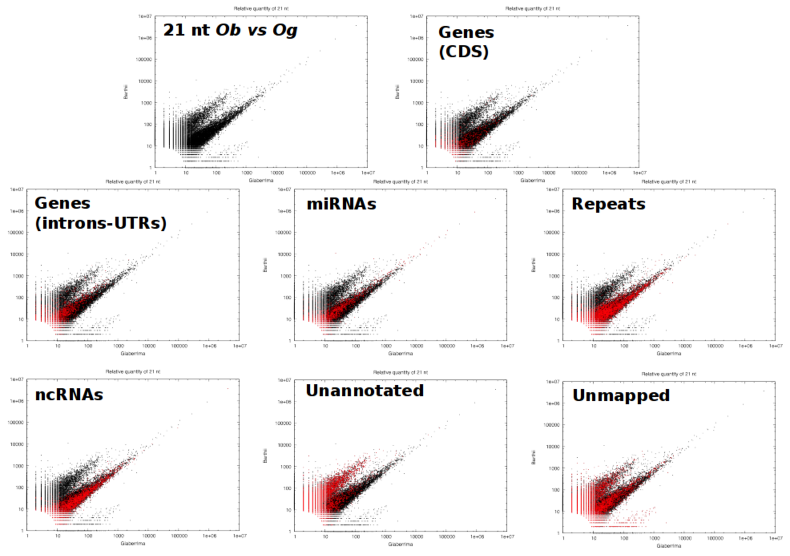
- Level of expression in different conditions or in different individuals
- Variation in sequences
- Variation in specific forms
- Detection of new forms

Example in smallRNA Transcriptomics



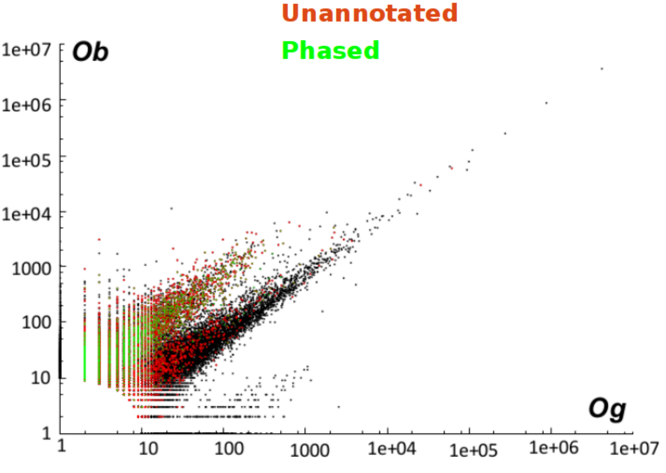
From Ta et al, 2015

Example in smallRNA Transcriptomics



From Ta et al, 2015

Example in smallRNA Transcriptomics



From Ta et al, 2015

Diagnostic/Pathology

- Pre-diagnostic (Genetic illness, putative resistance)

Diagnostic/Pathology

- Pre-diagnostic (Genetic illness, putative resistance)
- Tumor sequencing

Diagnostic/Pathology

- Pre-diagnostic (Genetic illness, putative resistance)
- Tumor sequencing
- Viral sequencing

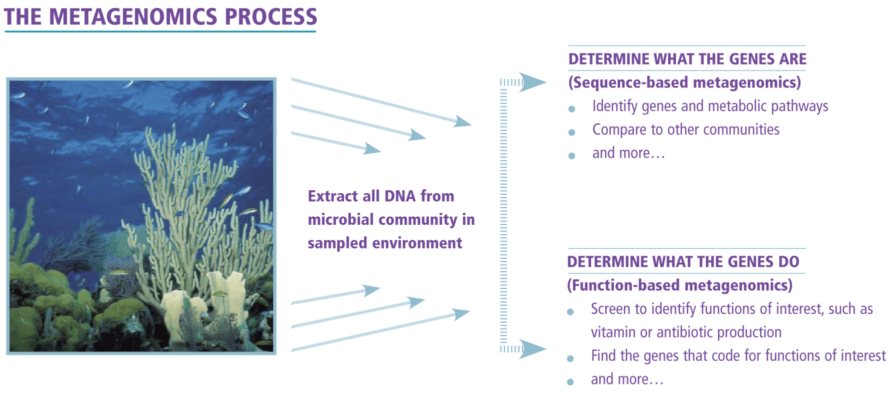
Diagnostic/Pathology

- Pre-diagnostic (Genetic illness, putative resistance)
- Tumor sequencing
- Viral sequencing
- Risk Assessement

Diagnostic/Pathology

- Pre-diagnostic (Genetic illness, putative resistance)
- Tumor sequencing
- Viral sequencing
- Risk Assessement
- Epidemiological Studies

Metagenomics at large



THE METAGENOMICS PROCESS

Extract all DNA from microbial community in sampled environment

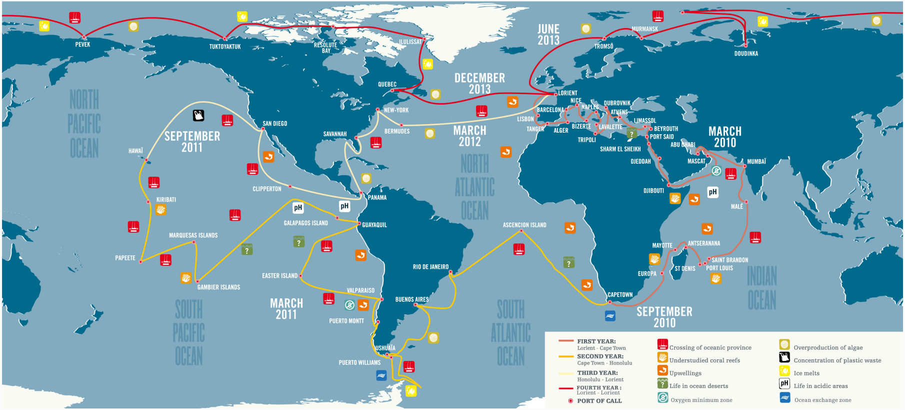
DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)

- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO
(Function-based metagenomics)

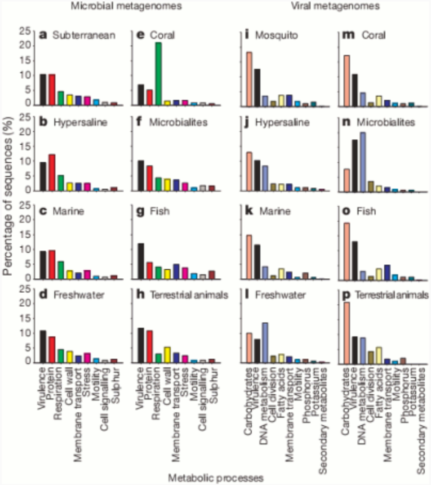
- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

Large Metagenomic assays



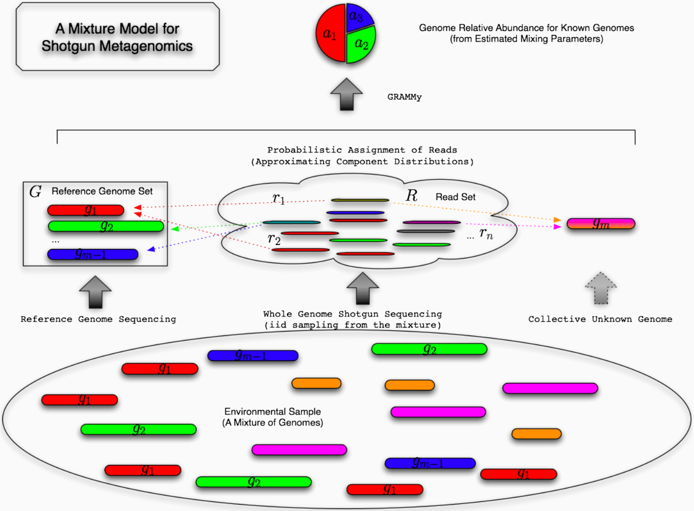
From Tara Ocean website

Functional Metagenomics

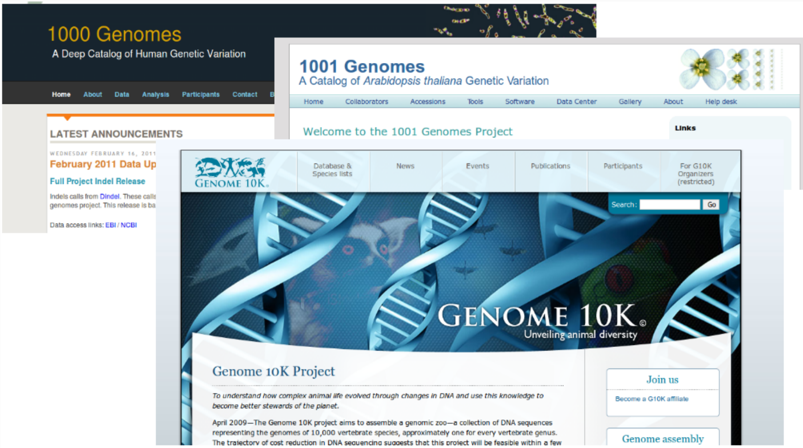


From Dinsdale et al, 2008

Barcoding


A Mixture Model for Shotgun Metagenomics
a_1
a_3
a_2
Genome Relative Abundance for Known Genomes (from Estimated Mixing Parameters)
GRAMMy
Probabilistic Assignment of Reads (Approximating Component Distributions)
G Reference Genome Set
g_1
g_2
...
g_{m-1}
r_1
R Read Set
r_2
... r_n
g_m
Reference Genome Sequencing
Whole Genome Shotgun Sequencing (iid sampling from the mixture)
Collective Unknown Genome
Environmental Sample (A Mixture of Genomes)
g_1
g_2
g_{m-1}

Large Projects



1000 Genomes

A Deep Catalog of Human Genetic Variation

LATEST ANNOUNCEMENTS

WEDNESDAY FEBRUARY 16, 2011

February 2011 Data Up

Full Project Indel Release

Indels calls from Dindel. These calls genomes project. This release is ba

Data access links: EBI / NCBI

1001 Genomes

A Catalog of *Arabidopsis thaliana* Genetic Variation

Welcome to the 1001 Genomes Project

Links

Genome 10K Project

To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

April 2009—The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The trajectory of cost reduction in DNA sequencing suggests that this project will be feasible within a few

Join us

Become a G10K affiliate

Genome assembly

Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics

Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014

Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015

Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available

Possibilities in the next 5-10 years (From a presentation in 2013)

- Real-time Transcriptomics
- Single-Cell Genomics -> DONE in 2014
- Single-Cells Transcriptomics (and smallRNA) -> DONE in 2015
- Personal Genomics medicine (ethical problems...) -> Available
- And any new ideas you will have...

This is the end...

- NGS technologies change the way of abording Biology

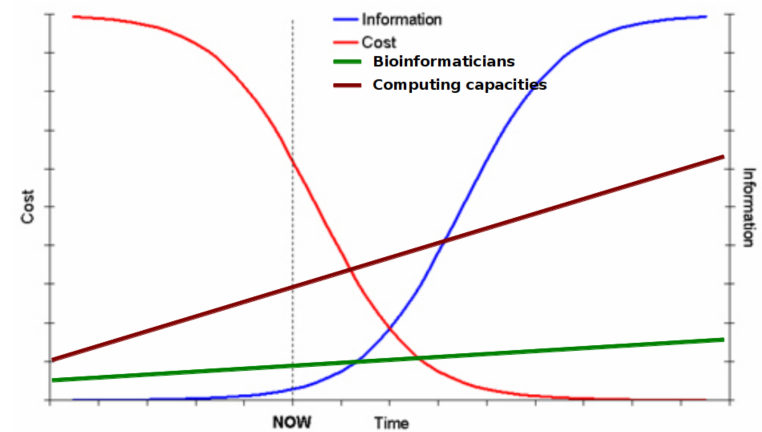
This is the end...

- NGS technologies change the way of abording Biology
- A lot of Possibilities, a lot of limits

This is the end...

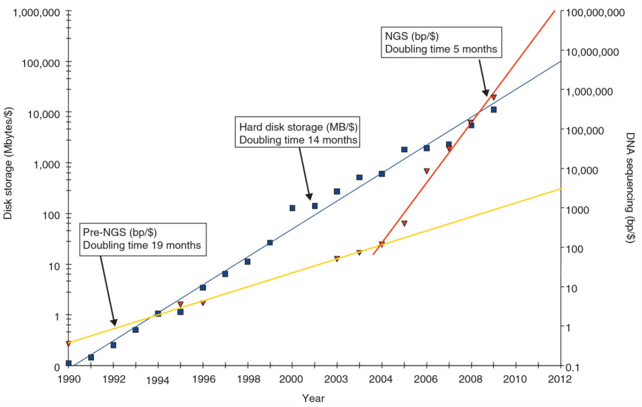
- NGS technologies change the way of abording Biology
- A lot of Possibilities, a lot of limits
- The main limit is no more Sequence, but Sample acquisition and Data treatment

Keep in mind!



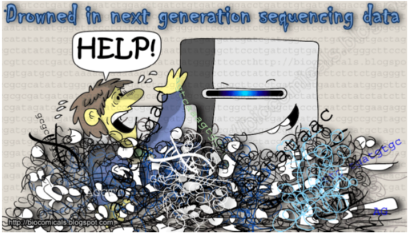
Information
Cost
Bioinformaticians
Computing capacities
Cost
Information
NOW
Time

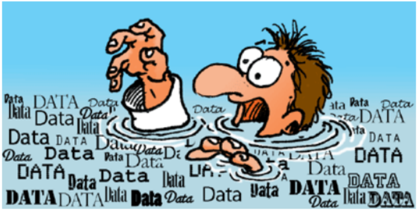
...From Data Rarity to Data Deluge



From L. Stein, 2010

Be Careful to data drowning!


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com



Thanks for your attention