Biodiversità Molecolare: aspetti di base e nuove tecnologie

Graziano PESOLE

Università di Bari & IBBE-CNR, Bari, Italy

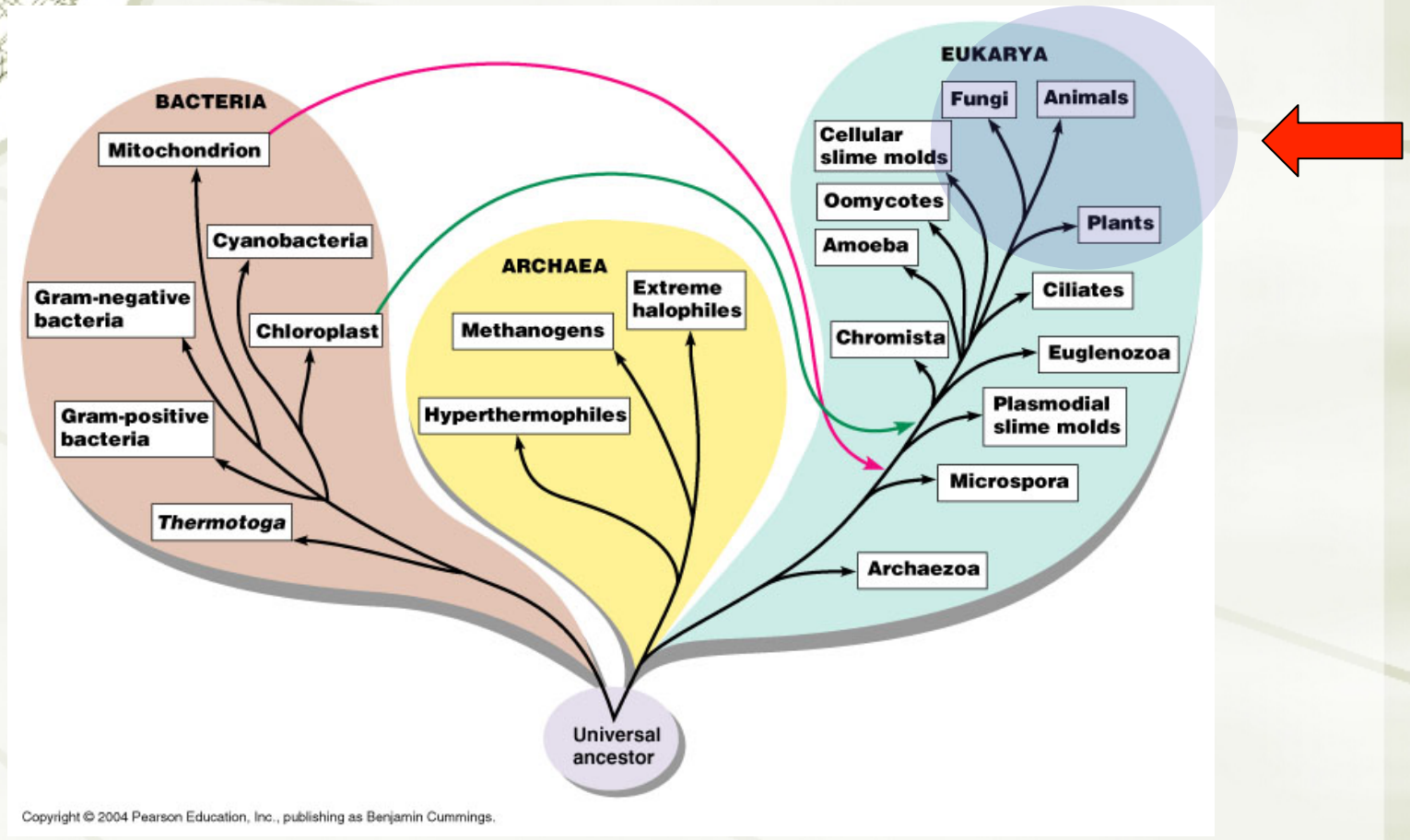
Seminari sulla Biodiversità
Bari, 18 Febbraio 2011

La Biodiversità

Per biodiversità si intende l'insieme di tutte le forme viventi, **geneticamente dissimili**, e degli ecosistemi ad esse correlati. Quindi biodiversità implica tutta la variabilità biologica: di geni, specie, habitat ed ecosistemi.

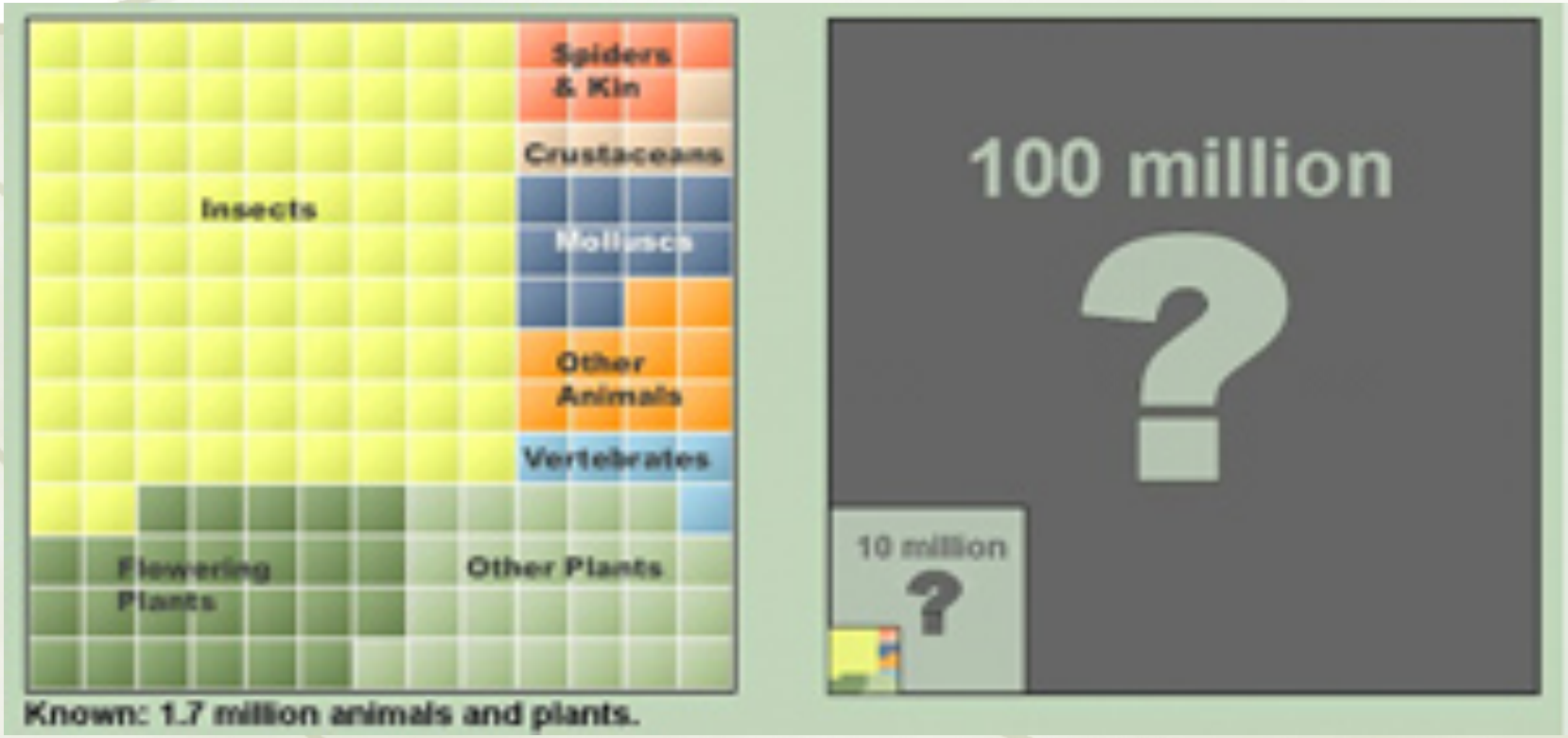
La Biodiversità

La gran parte della Biodiversità del pianeta non è visibile, ma è costituita da organismi microscopici (batteri, virus e eucarioti unicellulari).



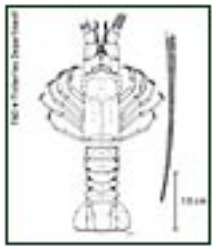
Quante sono le specie viventi?

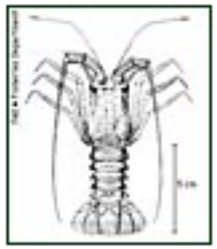
Quante sono le specie viventi? Oggi conosciamo circa 1,7 milioni di specie di animali e piante (senza considerare i microbi). Tuttavia, non conosciamo ancora il numero di specie diverse che popola la terra.. che potrebbero arrivare sino a 100 milioni.



Come riconosciamo una specie?

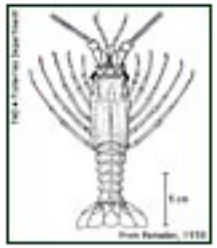
Normalmente, le specie, almeno nell'ambito di animali e piante possono essere riconosciute (e quindi classificate) in base alle proprietà morfologiche.

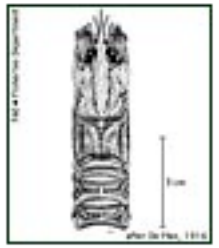

African spear lobster


American blunthorn lobster


American lobster


Andaman lobster


Arabian whip lobster


Arafura lobster

Come riconosciamo una specie?

Tuttavia si può osservare una grande varietà morfologica, nell'ambito di una stessa specie

Come riconosciamo una specie?

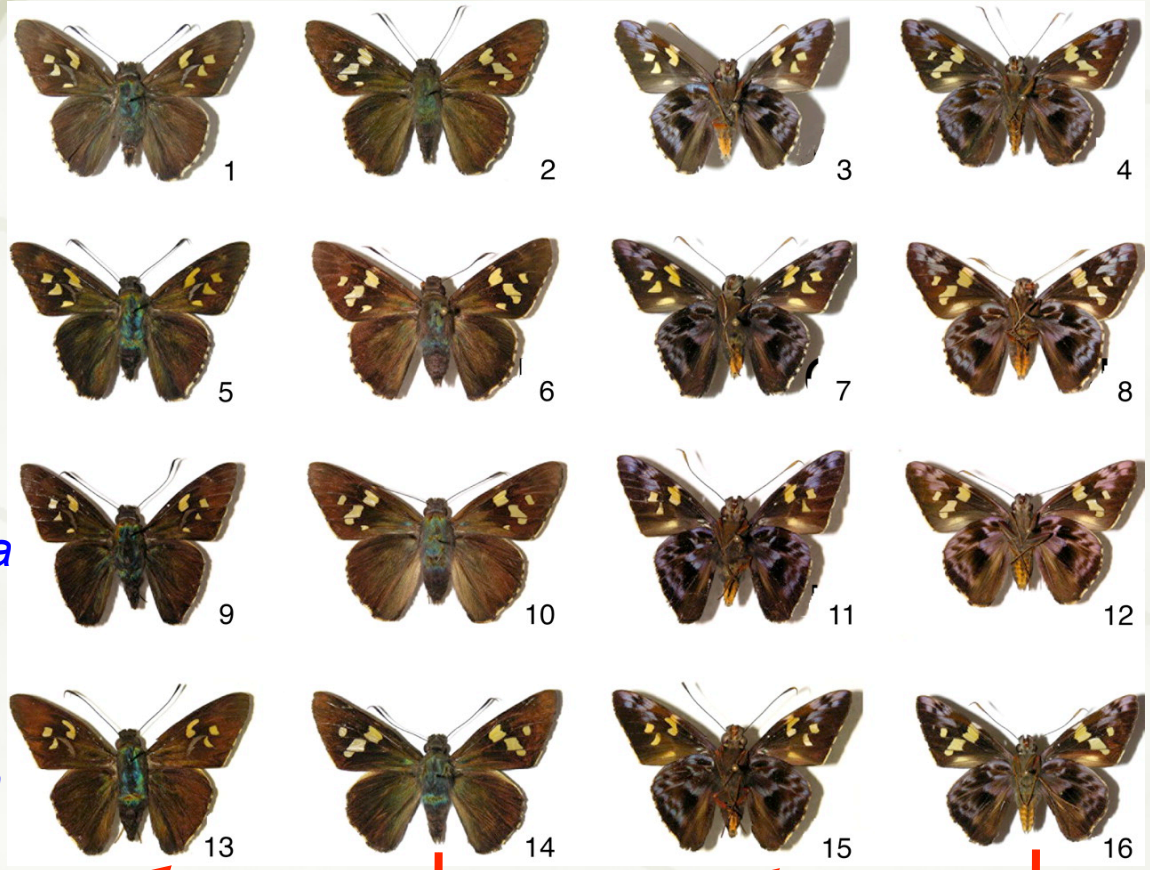
… e specie differenti possono apparire morfologicamente identiche (es. *Perichares sp.*)

P. adela

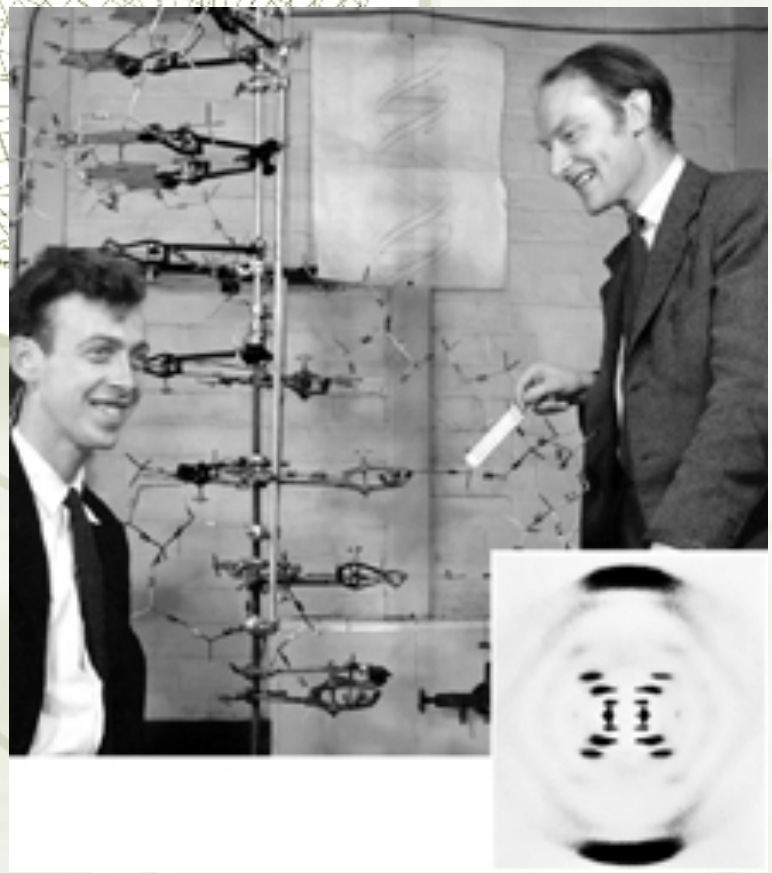
P. poaceaphaga

P. geonomaphaga

P. prestoeaphaga



Il test del DNA per l'identificazione delle specie?



L'individualità di ogni organismo vivente è racchiuso nel suo patrimonio genetico, costituito dal DNA.

In particolare, l'informazione genetica specifica di ogni organismo è costituita dalla sequenza delle 4 basi (**A, C, G, T**).

La sequenza delle basi nel genoma è quasi identica nell'ambito della stessa specie (>95%). Le piccole differenze intra-specifica sono quelle che consentono alla Polizia Scientifica di individuare gli autori dei crimini da piccolissime tracce di materiale biologico.

Il test del DNA per l'identificazione delle specie?

Il patrimonio genetico di specie differenti presenza delle differenze nella sequenza genomica in misura proporzionale alla divergenza evolutiva tra le specie, ma che comunque consente di identificare le specie in modo non ambiguo. Costituisce quindi una sorta di **carta d'identità genetica** della specie.

L'analisi genetica consente anche di identificare le specie conosciute a prescindere dalle loro caratteristiche morfologiche (es. batteri o altri organismi microbici) ma anche nuove specie sinora sconosciute.

Due Specie criptiche in *C. intestinalis*

L'analisi comparativa dei genomi mitocondriali ha permesso di identificare in modo incontestabile l'esistenza di due specie criptiche in *C. intestinalis* (Iannelli et al., Trends in Genetics, 2007, 23:419-22)

Quattro distinte linee di evidenza osservate analizzando le caratteristiche del mtDNA risultano non compatibili con la variabilità intra-specie:

1. **Ordine genico**
2. **Numero/dimensioni delle regioni non codificanti**
3. **Composizione in basi, e**
4. **Divergenza tra le sequenze**

Queste conclusioni sono anche supportate da uno studio basato sull'analisi degli ibridi derivati da incroci tra Cione di tipo A e tipo B, che risultano sterili a causa di gametogenesi difettiva (Caputi et al. PNAS, 2007, 104:9364-9).

Il nostro approccio è basato su un'analisi comparativa su larga scala del genoma mitocondriale rappresenta uno strumento estremamente efficiente e affidabile per l'identificazione di specie criptiche.

Sequenziamento del DNA

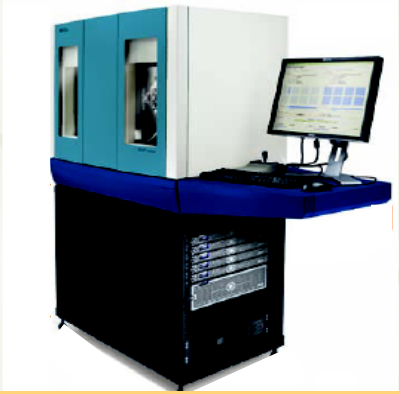
Per determinare la carta d'identità genetica di un organismo (genoma) o di un insieme di organismi (metagenoma) è quindi necessario determinare la sequenza di tutte le molecole di DNA presenti nel campione.

.. AGGCTACTAATTCGATTAGAGTTAGGGAGAGGGGGAGT..

E' attualmente in corso una grande rivoluzione tecnologica che ha portato allo sviluppo di strumenti (**next-generation sequencing**) che consentono di determinare la sequenza del DNA con elevatissima efficienza e costi estremamente contenuti.

Roche / 454 Genome Sequencer FLX titanium (400 bp, 400 Mb / run)

Illumina / Solexa Genetic Analyzer HiSeq (50-100 bp, 200 Gb / run)



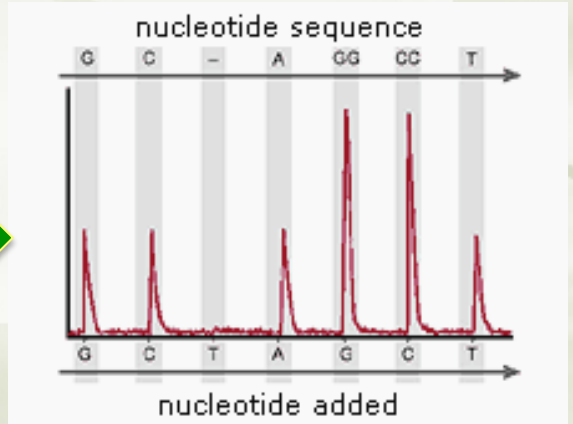
Applied Biosystems SOLiD 4 (50 bp, 100 Gb / run)

454 Pyrosequencing

$(DNA)_n$ + dNTP → DNA polimerasi → PPi + $(DNA)_{n+1}$

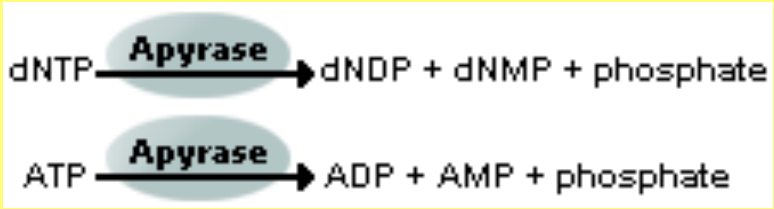
PPi + APS → Solforilasi → ATP + Si

ATP + Si + luciferina + O_2 → Luciferasi → AMP+PPi+ossiluciferina +CO_2 + luce



454 Pyrosequencing

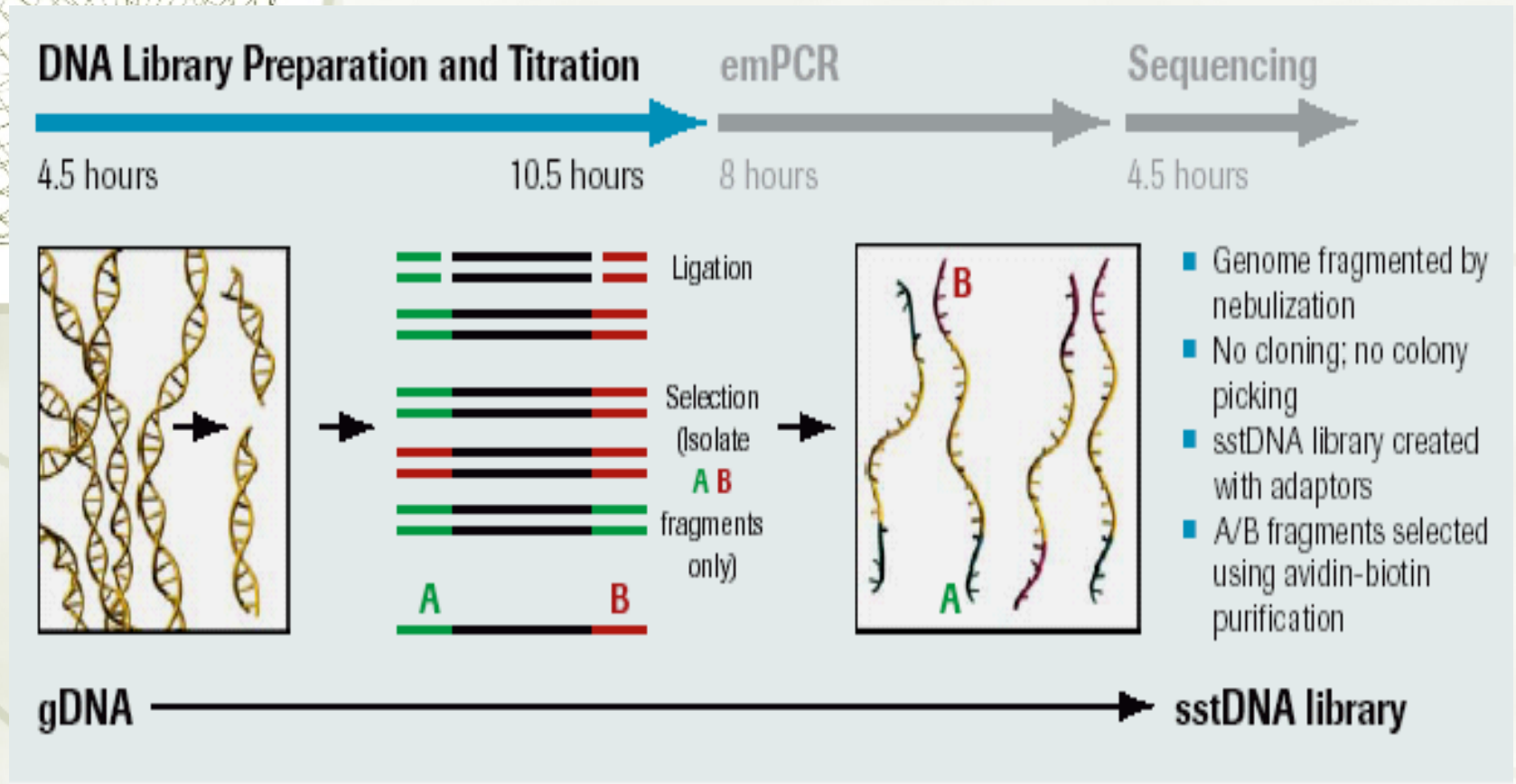
Al termine di ogni ciclo i nuleotidi in eccesso sono degradati dall'enzima Apirasi



Il dATP è un substrato naturale della luciferasi. Per questo viene utilizzato per la polimerizzazione un analogo dell'ATP (deoxy-adenosine α-tio-triphosphate (dATPS), che viene efficientemente incorporato dalla DNA polimerasi, ma non è un substrato della luciferasi.

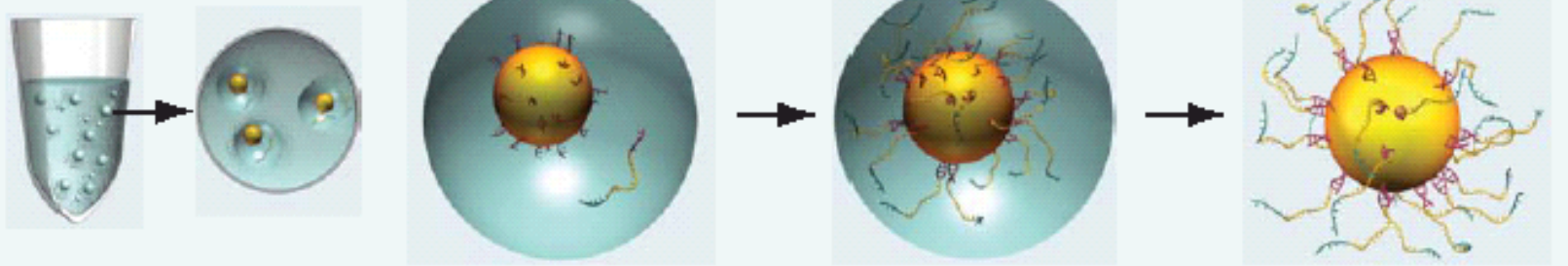
454 Pyrosequencing

Phase I



454 Pyrosequencing

Phase II



Anneal sstDNA to an excess of DNA Capture Beads

Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors, enrich for DNA-positive beads

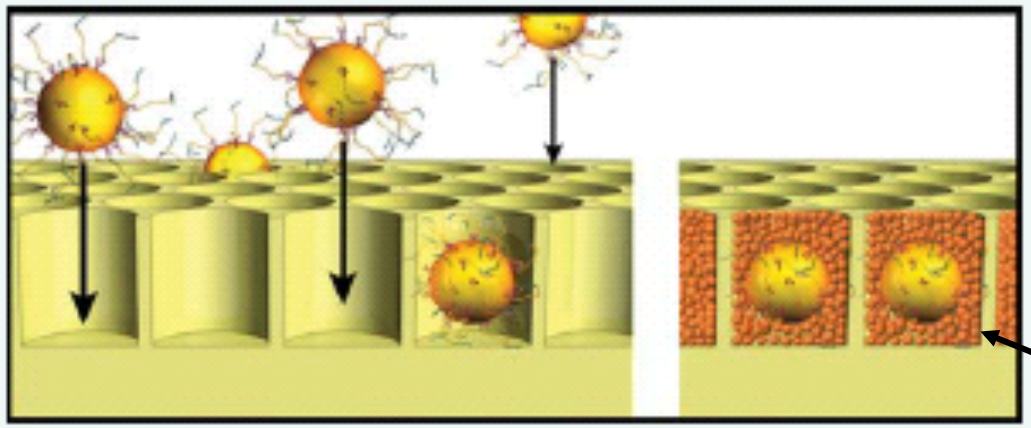
sstDNA library → Clonally-amplified sstDNA attached to bead

454 Pyrosequencing

Phase III

DNA Library Preparation and Titration → emPCR → Sequencing

4.5 hours — 10.5 hours | 8 hours | 4.5 hours



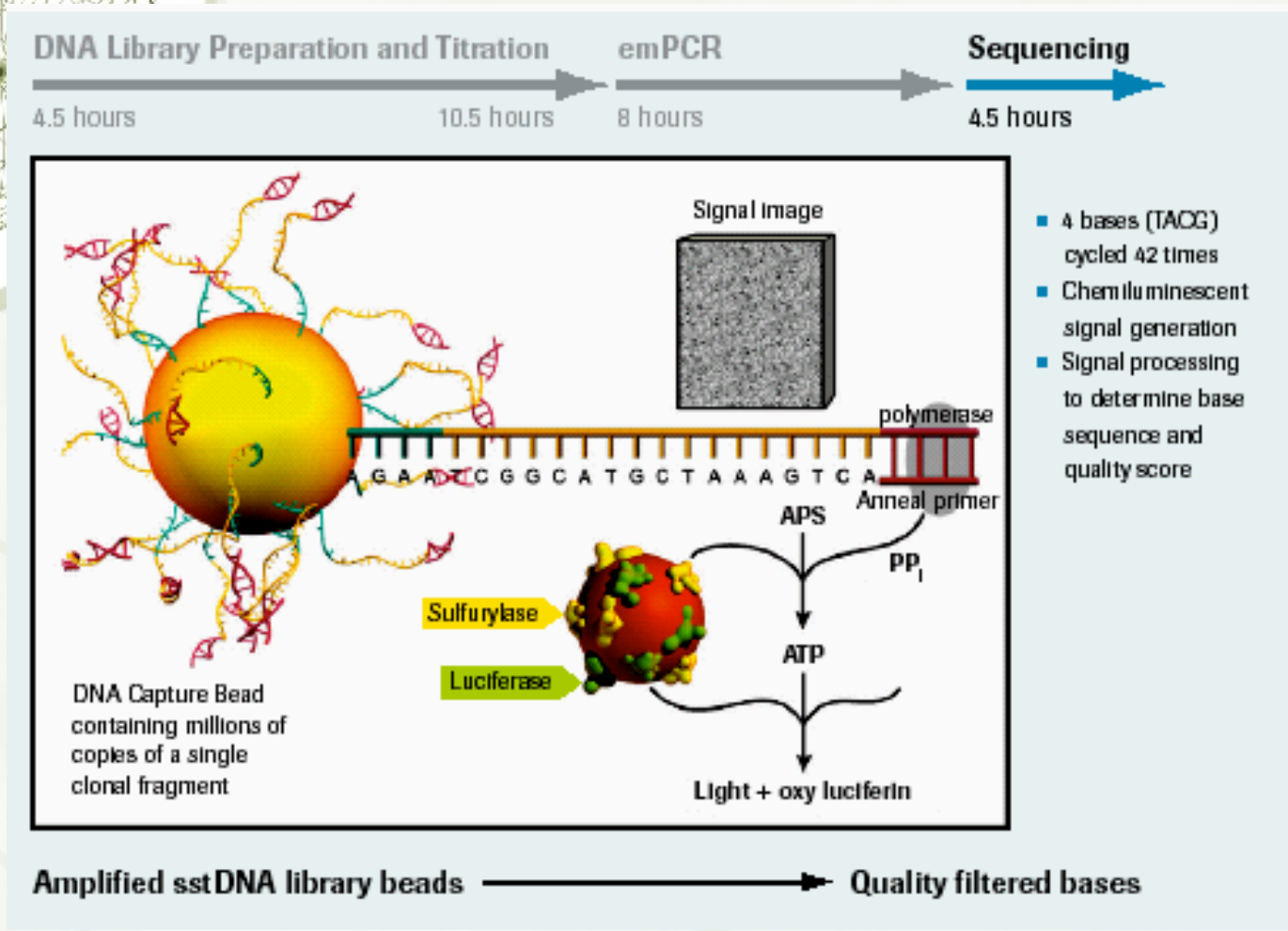
- Well diameter: average of 44µm
- 200,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

Pyrosequencing Enzyme beads

Amplified sstDNA library beads → Quality filtered bases

454 Pyrosequencing

Phase IV

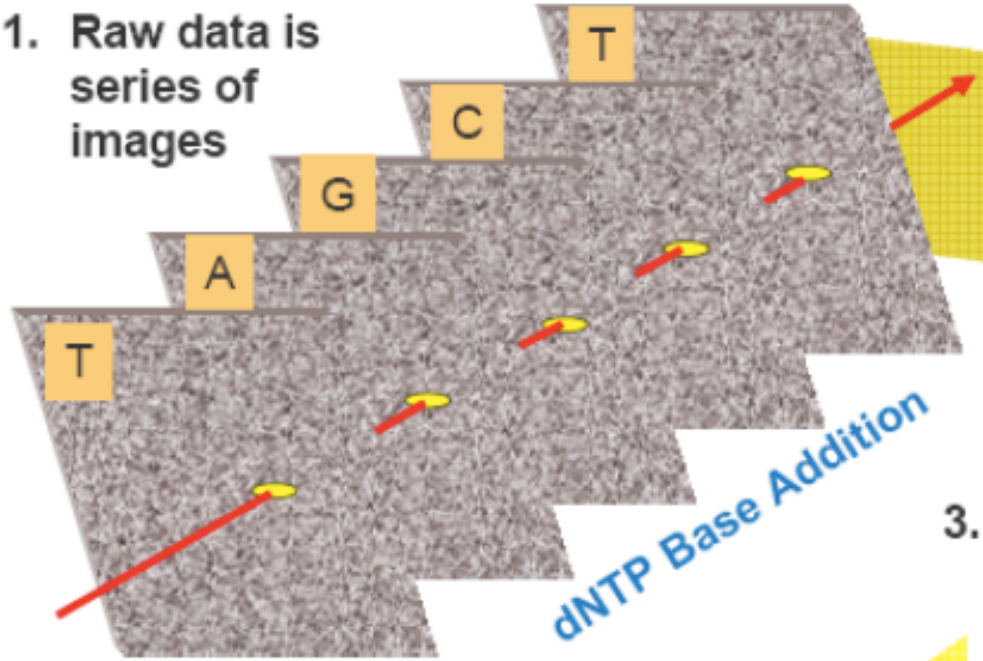


454 Pyrosequencing

GS FLX Data Analysis

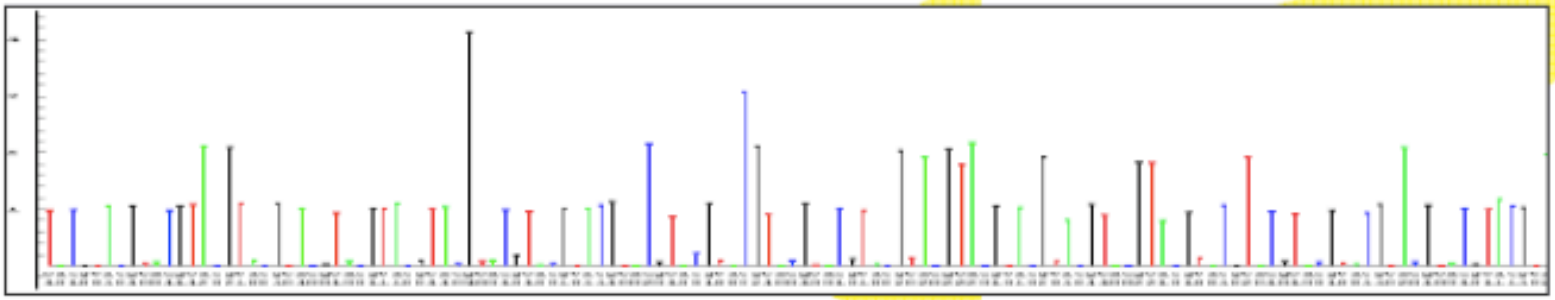
Image Processing

1. Raw data is series of images



2. Each well's data is extracted, quantized and normalized

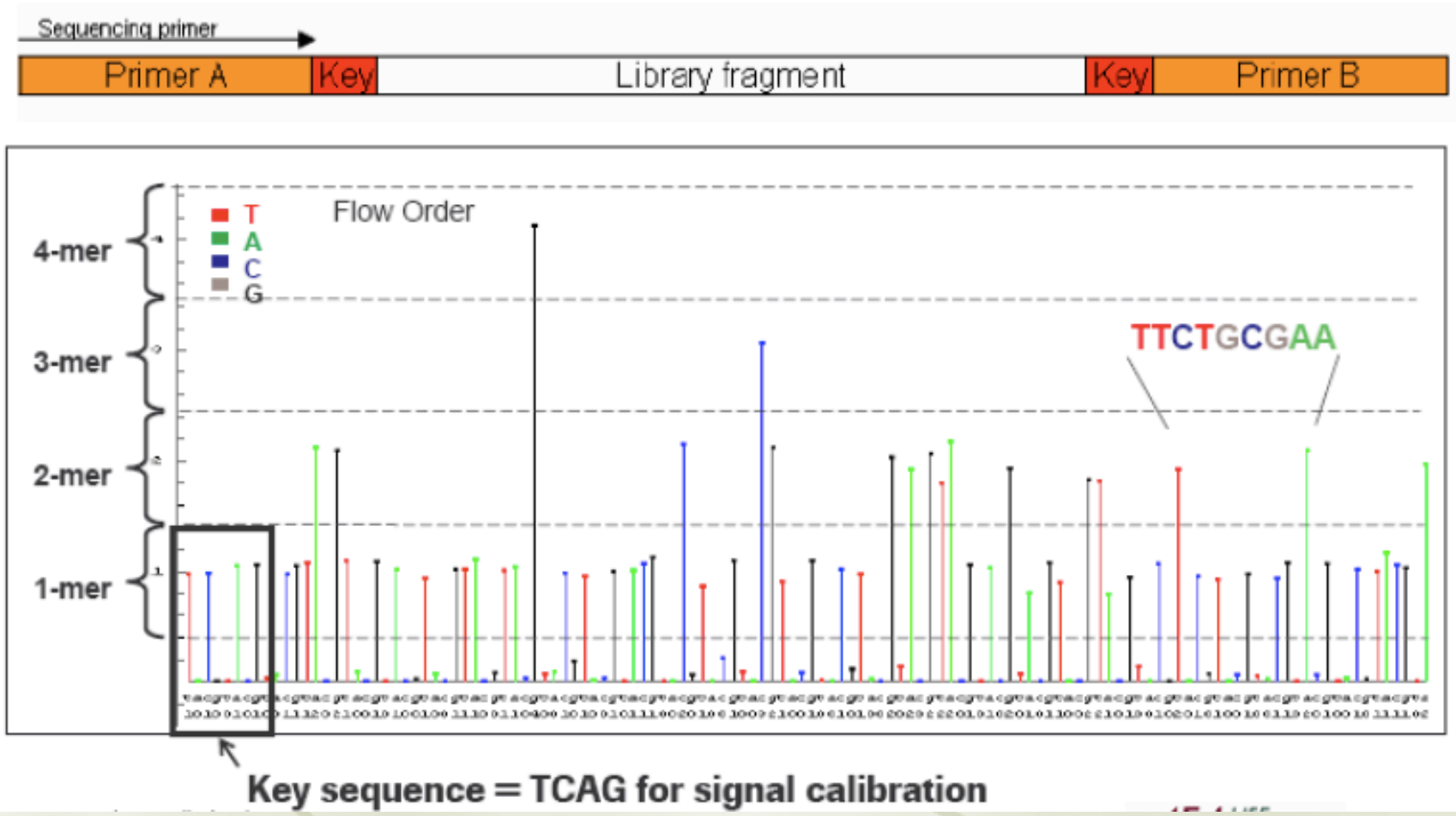
3. Read data converted into "flowgrams"

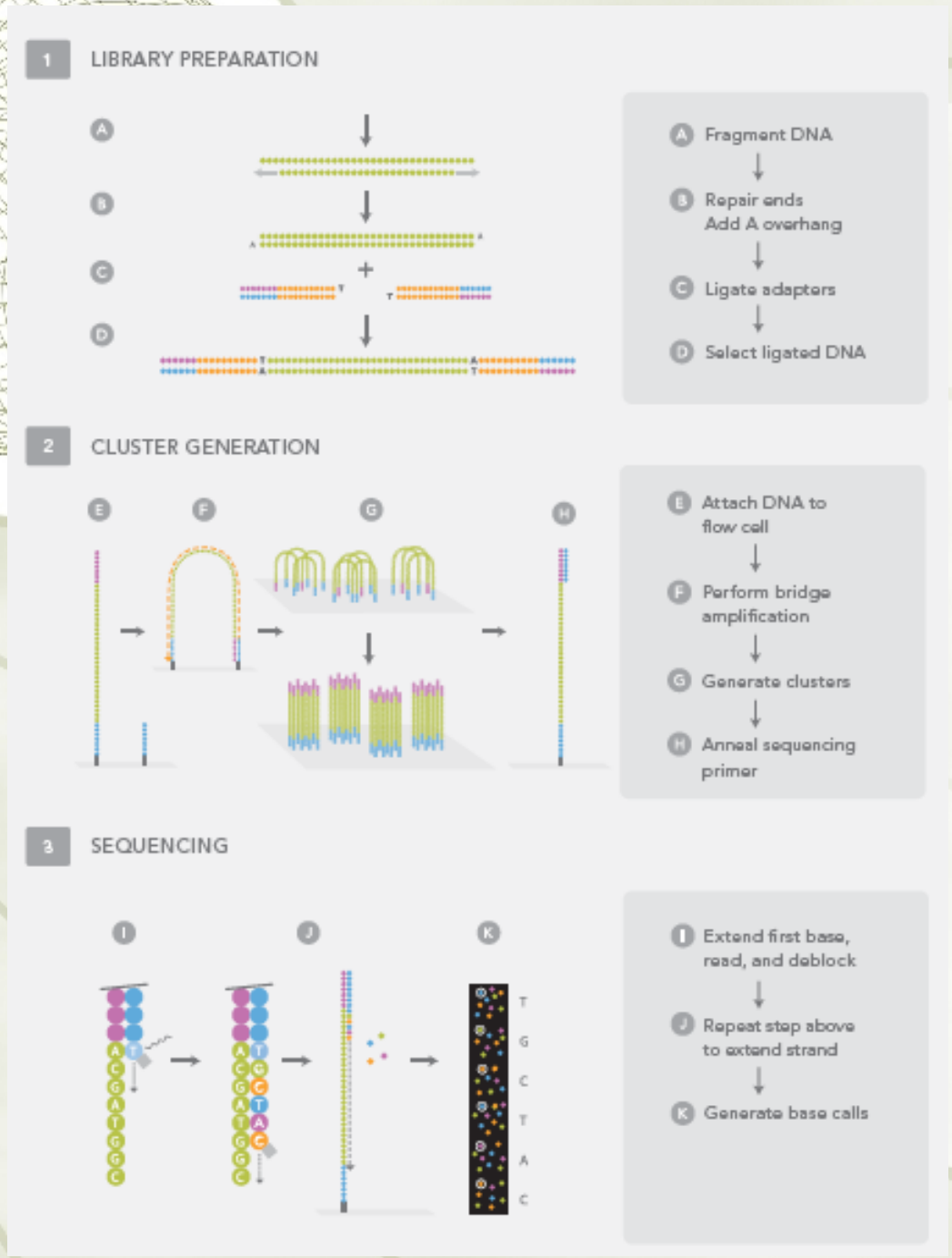


454 Pyrosequencing

GS FLX Data Analysis

Flowgram Generation

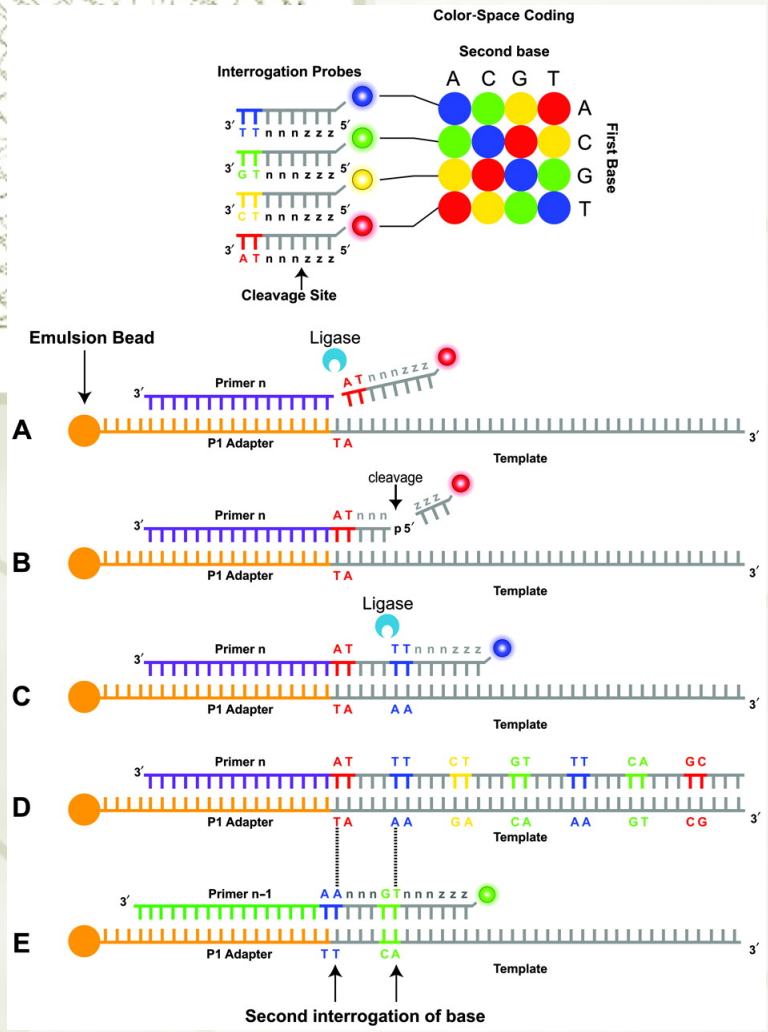




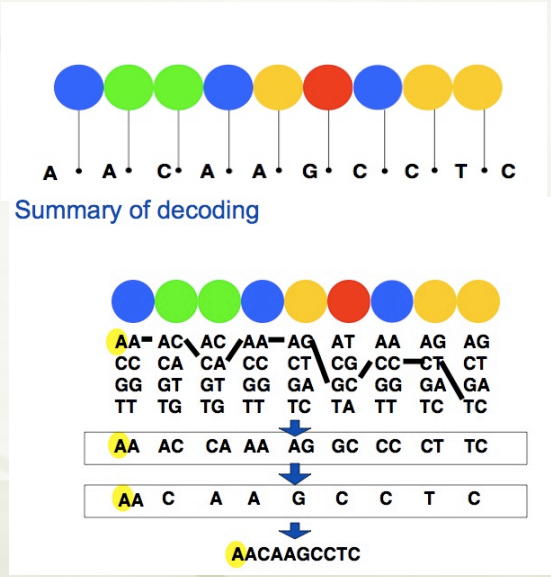
Illumina Sequencing by synthesis

Il sistema Illumina effettua una amplificazione “a ponte” (bridge amplification) dei frammenti su un vetrino, seguita dal sequenzimento associato alla sintesi del DNA (come nel metodo di Sanger) utilizzando terminatori reversibili.

Applied Biosystems sequencing by ligation



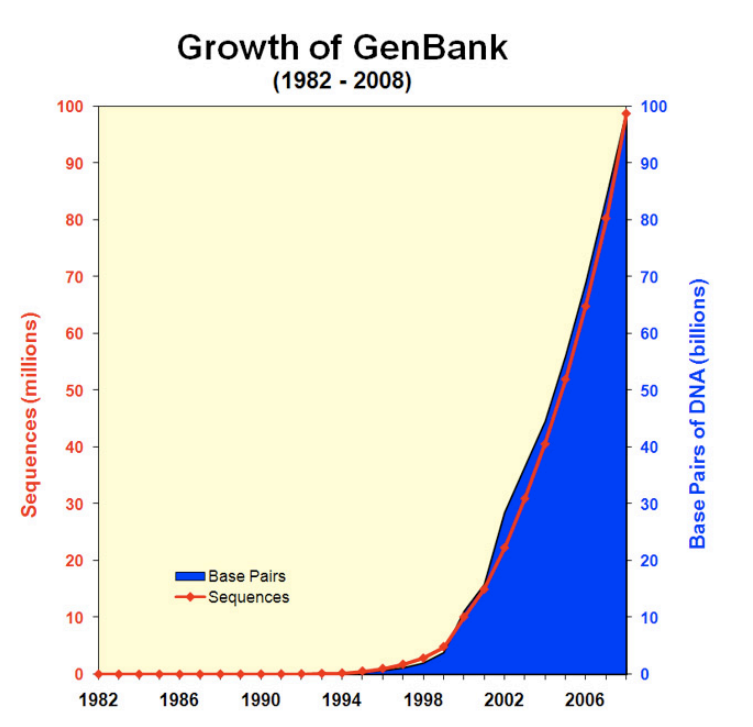
Il sistema SOLID di AB segue una procedura simile, nelle prime fasi al sequenziamento 454 (frammentazione, aggiunta di adattatori, amplificazione in emulsione). Il sequenziamento è basato sulla ligazione progressiva di oligonucleotidi in cui un diucleotide ha sequenza nota.
Il risultato è una sequenza di "colori" ciascuno dei quali rappresenta uno specifico dinucleotide.



Voelkerding, K. V. et al. Clin Chem 2009;55:641-658

Next-Generation Sequencing Statistics

	Sanger	Roche	Illumina	ABI Solid
Read (L)	1000	400	100 (x2)	75 (x2)
Resa (Gbp)	0.0001	0.4	200	300



Una singola piattaforma di sequenziamento è oggi in grado di produrre più dati di quanti ne siano stato prodotti negli ultimi 30 anni.

La rivoluzione NGS per lo studio della Biodiversità

La rivoluzione NGS rende oggi possibile intraprendere analisi genomiche su una dimensione di scala fino a poco tempo fa inimmaginabile. Allargheremo così in modo incredibile le attuali conoscenze sulla Biodiversità. E' prevedibile che un numero enorme di nuove specie venga identificato nel prossimo futuro.

Applicazioni delle piattaforme NGS

- **DNA**
  - genome resequencing (analisi SNPs, GWAS)
  - sequencing de novo
  - identificazione di varianti strutturali del genoma (cancer genome)
  - Epigenomica (stato della cromatina e siti di metilazione)
  - **Metagenomica (analisi tassonomica/funzionale dei campioni ambientali)**

- **RNA**
  - Analisi qualitativa e quantitativa del trascrittoma
  - Identificazione di miRNA e altri ncRNA
  - RNA editing
  - **Metatrascrittomica (analisi funzionale dei campioni ambientali)**

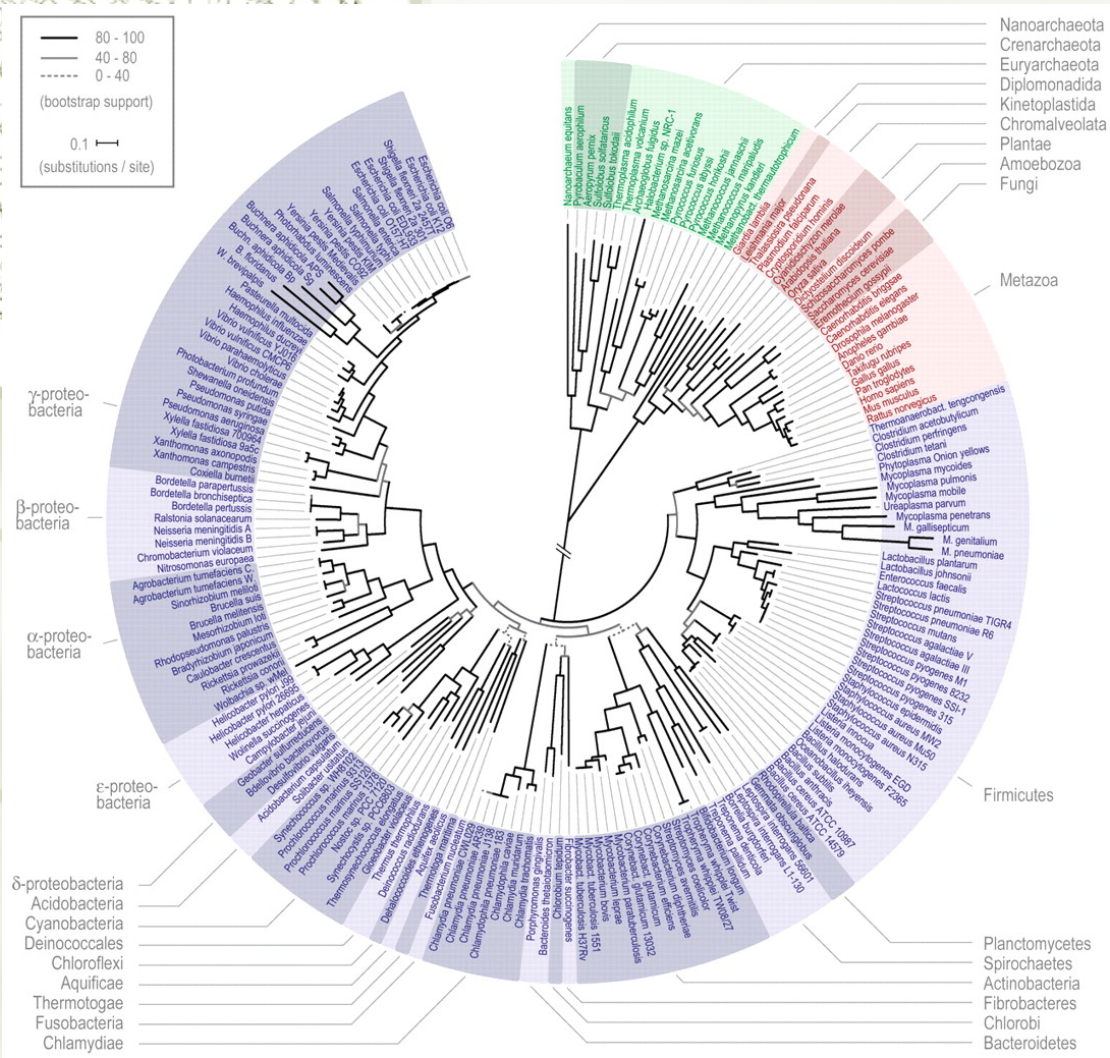
Metagenomica: definizione e obiettivi

Se consideriamo un certo ambiente (es. acqua, suolo, etc.), esso sarà popolato da una comunità più o meno complessa di (micro)organismi.

- L'insieme del materiale genetico presente in un campione ambientale, definito **metagenoma,** sarà quindi rappresentativo delle specie che lo popolano. Analogamente, il **metatrascrittoma** sarà rappresentativo delle funzioni espresse in quel determinato ambiente.

- Il principio base della Metagenomica è quindi che la **Biodiversità** di qualunque ambiente è pienamente rappresentata dal materiale genetico in esso presente.

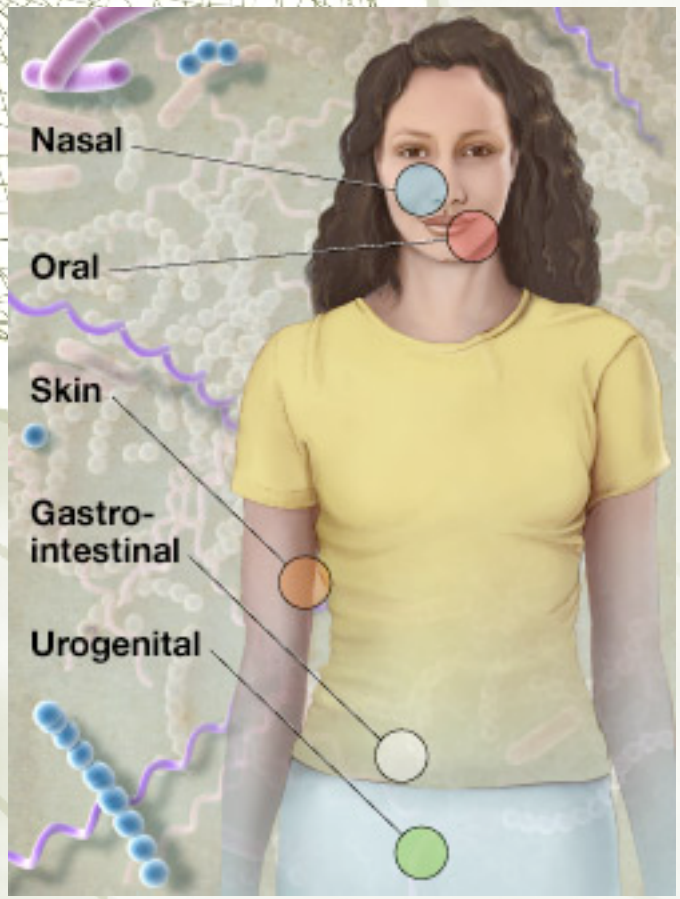
Metagenomica: esplorazione dell'evoluzione



L'esplorazione su larga scala dei dati metagenomici ci offre la straordinaria opportunità, prima impensabile, di far luce sulla complessità tassonomica di tutti i viventi … e soprattutto di ottenere una panoramica completa dei prodotti e dei processi dell'evoluzione in una grande varietà di ambienti e condizioni.

Sarà possibile scoprire nuovi geni e nuove funzioni per un largo spettro di processi e applicazioni biotecnologiche.

Il Microbioma umano



Anche l'uomo è un ambiente estremamente complesso. Il nostro corpo è costituito da circa 10^{13} cellule, ma contiene un numero dieci volte maggiore (circa 10^{14}) di cellule batteriche. Il cosiddetto "microbioma umano" ha una profonda influenza sulla fisiologia dell'organismo, sulla nutrizione, e risulta cruciale per la nostra salute. Difatti, esso fornisce nutrienti e vitamine, coadiuva la risposta alle infezioni e la detossificazione da diverse sostanze tossiche.

La Metagenomica ora rende possibile la caratterizzazione della composizione e della dinamica di popolazione della comunità microbica umana, e le interazioni cooperative (o antagoniste) con le cellule e i tessuti umani.

Metagenomica: strategie di analisi

Targeted-oriented metagenomics

Il sequenziamento massivo in parallelo di una specifica regione target (es. 16S rRNA o ITS nei batteri) di ampliconi ottenuti utilizzando primer universali specifici per un determinato (il più vasto possibile) gruppo tassonomico. Una regione di 648 bp del gene mitocondrale per la citocromo c ossidasi (cox1) è stato proposto come **barcode genetico** dei Metazoi.

Shotgun metagenomics

Sequenziamento shorgun del DNA (o RNA) estratto da campioni ambientali.

Il primo approccio è particolarmente adatto per specifici gruppi tassonomici, per cui sono disponibili primer universali che siano in grado di amplificare una specifica regione genomica per un gran numero di specie (es. batteri).
Il sequenziamento massivo dell'RNA (Metatrascrittomica) può offrire un quadro completo del profilo di espressione della comunità microbica all'interno del campione ambientale il esame, in termini qualitativi e quantitativi.

Metagenomica: strategie di analisi

Targeted-oriented metagenomics

Identificazione a livello di specie o gruppo tassonomico

Shotgun metagenomics

Caratterizzazione a livello di specie (taxon), gene, o pathway

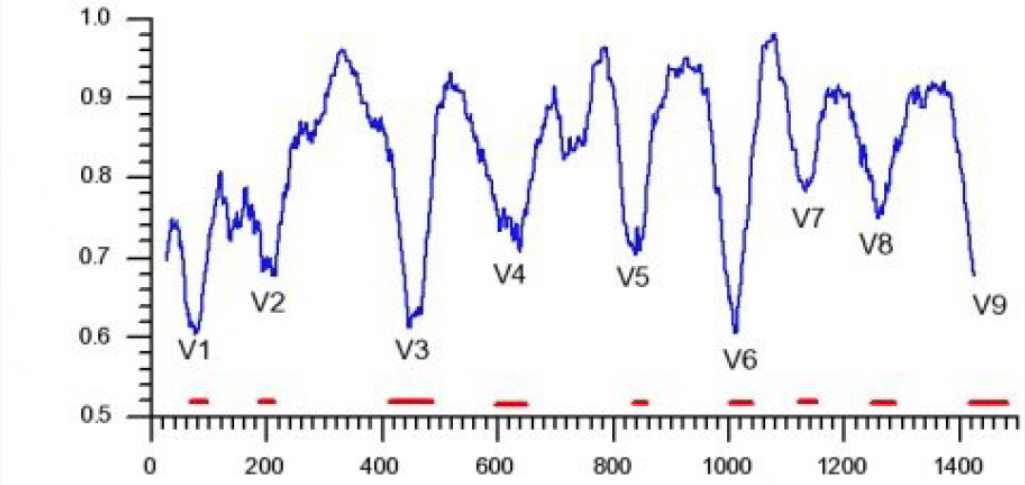
Targeted-oriented Metagenomics

Bisogna definire una specifica regione genomica che è ubiquitariamente presente nel gruppo tassonomico in esame, e che abbia ben determinate caratteristiche:

- Regioni fiancheggianti altamente conservate, che consentano l'amplificazione e il sequenziamento in in ampio spettro di specie mediante primer universali.

- Sufficiente variabilità della regione interna per poter facilmente discriminare tra differenze intra-specifiche e inter-specifiche

Small subunit 16S ribosomal RNA structure



Le regioni variabili dei 16S rRNA batterici (V1-V9) vengono generalmente utilizzate per analizzare i batteri, cox1 per i metazoi, e rbcl o matK per le piante.

La Bioinformatica

La grande quantità di dati molecolari (principalmente sequenze di acidi nucleici) prodotti nell'ambito dei progetti di metagenomica richiede strumenti adeguati per il loro immagazzinamento, organizzazione ed analisi. Di tutto questo si occupa la **Bioinformatica**, una nuova disciplina che ormai ha assunto un ruolo di primo piano nella ricerca e sviluppo in una grande varietà di settori biotecnologici, da quello biomedico a quello ambientale e agro-alimentare.

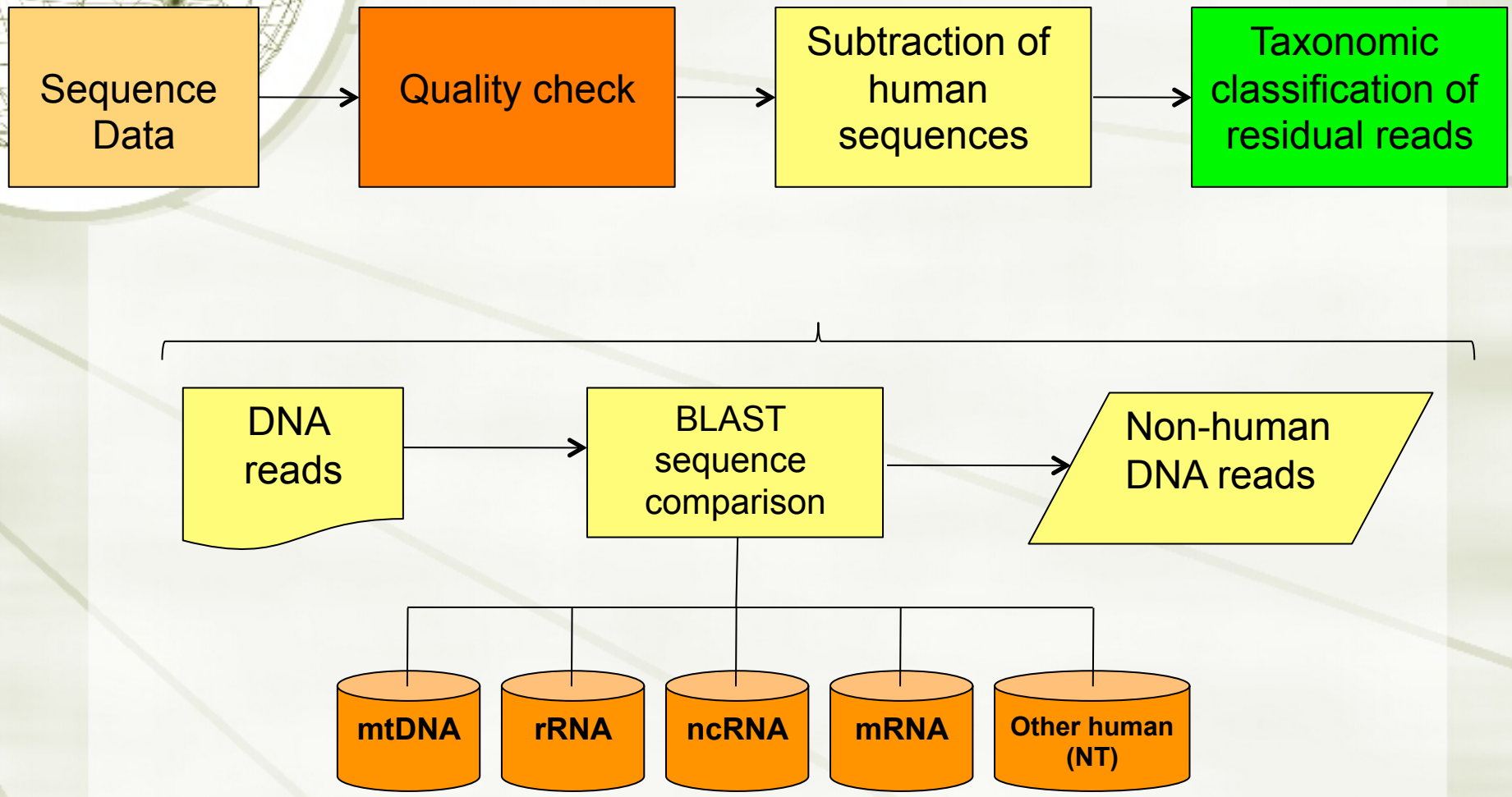
La Bioinformatica richiede il concorso di competenze multidisciplinari, particolarmente quelle in ambito biomolecolare e informatico.

LIFEWATCH

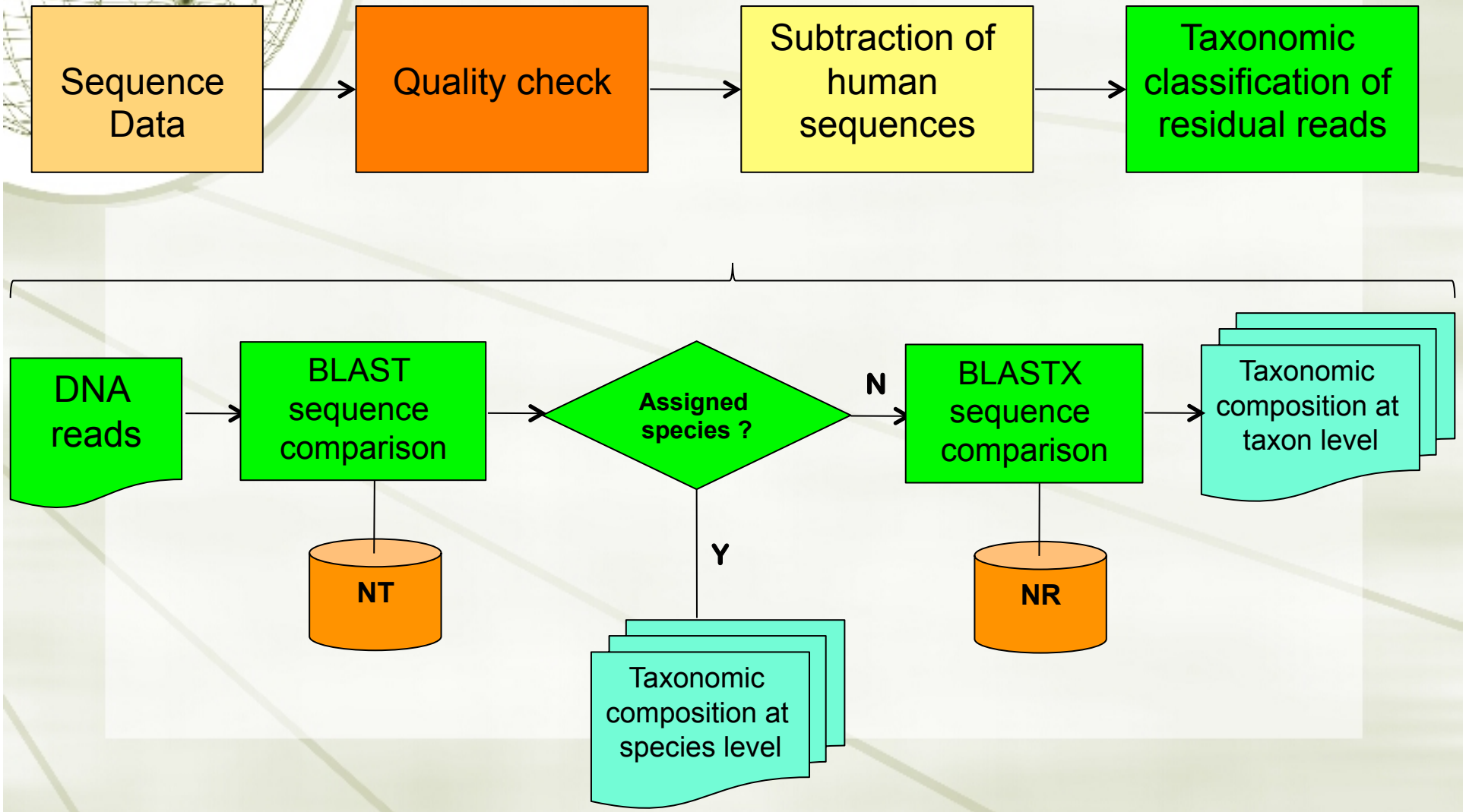
Research infrastructure for biodiversity and ecosystem research

www.lifewatch.eu

Workflow di analisi bioinformatica per l'analisi di un campione clinico



Workflow di analisi bioinformatica per l'analisi di un campione clinico



Classificazione tassonomica

Assegnazione di Specie

Una specie può essere assegnata ad una determinata sequenza (l'osservazione della sequenza è indicativa della presenza della specie) se la divergenza osservata con la sequenza target è compatibile con la variazione intra-specifica e con il tasso di errore nel sequenziamento (%identità ≥95%), e se l'allineamento non viene osservato per caso (P-value e lunghezza).

L'informazione tassonomica relativa alla sequenza target è fornita dalla banca dati **NCBI Taxonomy** (es. per l'uomo, TaxID=9606). I TaxID sono assegnati anche a livelli tassonomici superiori (genere, famiglia, etc.) e inferiori (sottospecie, ceppo, ecc.) rispetto alla specie stessa.

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	16,894,613	16,894,588
Nucleotide EST	8,301,471	8,301,471
Nucleotide GSS	1,293,831	1,292,505
Protein	529,402	529,306
Structure	15,602	15,602

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	96	113	525	143	877
Bacteria	1109	1995	15803	7604	26511
Eukaryota	16453	50641	197126	16598	280817
Fungi	1193	3613	21327	1372	27504
Metazoa	12045	31661	87582	8000	139288
Viridiplantae	1964	13198	81195	6204	102561
Viruses	500	342	6655	50851	58348
All taxa	18183	53099	225433	75231	371945

Assegnazione di Specie

Può accadere che la sequenza in esame mappi più o meno ugualmente bene con più sequenze della banca dati, appartenenti a specie diverse.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
XM_002807888.1	PREDICTED: Callithrix jacchus E3 ubiquitin-protein ligase HUWE1-like (LOC100405596),	346	346	52%	5e-92	99%
XR_005443.2	PREDICTED: Rattus norvegicus hypothetical protein LOC501546 (LOC501546), miscRNA	346	346	52%	5e-92	99%
XR_009642.2	PREDICTED: Rattus norvegicus hypothetical protein LOC501546, transcript variant 1 (LO	346	346	52%	5e-92	99%
XR_085367.1	PREDICTED: Oryctolagus cuniculus HECT, UBA and WWE domain containing 1 (LOC10034	346	346	52%	5e-92	99%
NM_031407.4	Homo sapiens HECT, UBA and WWE domain containing 1 (HUWE1), mRNA	346	346	52%	5e-92	99%
XM_001914731.1	PREDICTED: Equus caballus HECT, UBA and WWE domain containing 1 (HUWE1), mRNA	346	346	52%	5e-92	99%
NM_001110004.1	Bos taurus HECT, UBA and WWE domain containing 1 (HUWE1), mRNA	346	346	52%	5e-92	99%
XR_025627.1	PREDICTED: Pan troglodytes HECT, UBA and WWE domain containing 1 (HUWE1), mRNA	346	346	52%	5e-92	99%
BC001637.1	Homo sapiens cDNA clone IMAGE:3355758, **** WARNING: chimeric clone ****	346	346	52%	5e-92	99%
XM_001088879.1	PREDICTED: Macaca mulatta similar to HECT, UBA and WWE domain containing 1, transc	346	346	52%	5e-92	99%
XM_001089097.1	PREDICTED: Macaca mulatta similar to HECT, UBA and WWE domain containing 1, transc	346	346	52%	5e-92	99%
XM_001088987.1	PREDICTED: Macaca mulatta similar to HECT, UBA and WWE domain containing 1, transc	346	346	52%	5e-92	99%
BC014208.1	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1	346	346	52%	5e-92	99%
XM_859412.1	PREDICTED: Canis familiaris similar to HECT, UBA and WWE domain containing 1, transcr	346	346	52%	5e-92	99%
XM_859397.1	PREDICTED: Canis familiaris similar to HECT, UBA and WWE domain containing 1, transcr	346	346	52%	5e-92	99%
XM_538052.2	PREDICTED: Canis familiaris similar to HECT, UBA and WWE domain containing 1, transcr	346	346	52%	5e-92	99%
DQ097177.1	Homo sapiens Mcl-1 ubiquitin ligase (MULE) mRNA, complete cds	346	346	52%	5e-92	99%
AY772009.1	Homo sapiens ARF-binding protein 1 mRNA, complete cds	346	346	52%	5e-92	99%
AB002310.2	Homo sapiens mRNA for KIAA0312 gene, partial cds	346	346	52%	5e-92	99%
AY929612.1	Homo sapiens LASU1 mRNA, partial cds	346	346	52%	5e-92	99%
AB071605.1	Homo sapiens mRNA for HECT domain protein LASU1, complete cds	346	346	52%	5e-92	99%
NM_021523.4	Mus musculus HECT, UBA and WWE domain containing 1 (Huwe1), mRNA	335	335	52%	1e-88	98%
AY772010.3	Mus musculus ARF-binding protein 1 mRNA, complete cds	335	335	52%	1e-88	98%
DQ097265.1	Mus musculus Mcl-1 ubiquitin ligase (Mule) mRNA, complete cds	335	335	52%	1e-88	98%

Nell'esempio specifico, la nostra sequenza si allinea abbastanza bene a 9 differenti specie (i.e. *Callithrix jacchus, Rattus norvegicus, Oryctolagus cuniculus, Homo sapiens, Equus caballus, Bos taurus, Pan troglodytes, Macaca mulatta, Canis familiaris*). In questo caso la specie non può essere assegnata univocamente, ma viene assegnata partizionandola su più specie in misura inversamente proporzionale alla significatività statistica dell'allineamento.

Dalla Biodiversità alle Biotecnologie

I milioni di specie che oggi popolano la terra, dagli organismi unicellulari più semplici come i batteri, fino a forme di vita multicellulari più complesse come l'uomo, sono il frutto dell'elaborazione di miliardi di anni di evoluzione in cui la selezione naturale ha operato in risposta a specifiche interazioni con l'ambiente (es. mutamenti climatici) e con le altre forme di vita che nel tempo hanno avuto origine e si sono estinte.

La decifrazione del patrimonio genetico degli organismi non solo ne facilita il riconoscimento ma ne svela anche il corredo di geni che li caratterizza, ciascuno dotato di specifiche funzioni, che sono il risultato del processo evolutivo che ha garantito la sopravvivenza dell'organismo nel suo ambiente.

Dalla Biodiversità alle Biotecnologie

Un gene, può essere considerato come l'unità elementare del patrimonio genetico di un organismo, in grado di guidare la sintesi di una (o più) proteina/e in grado di svolgere una specifica funzione.

I nuovi geni (o proteine) identificati nell'analisi metagenomica potrebbero trovare numerose applicazioni nel campo delle Biotecnologie.

Biotecnologie

Insieme di discipline che contempla l'utilizzo di organismi o di loro componenti cellulari nei processi industriali per la produzione di beni e servizi

Dalla Biodiversità alle Biotecnologie

L'esplorazione del metagenoma di un certo campione ambientale ci consente di condurre analisi su larga scala della complessa dinamica degli organismi presenti in un certo ecosistema, la cui sopravvivenza e adattamento dipende dalla mutua cooperazione di più specie. In questo modo si può anche far luce sul processo evolutivo che ha condotto alla biodiversità che oggi osserviamo e alle strategie che ne hanno consentito l'adattamento all'ambiente.

Ad esempio l'analisi metagenomica di campioni relativi ad ambienti estremi (freddo, caldo, arido) consente l'identifcazione di microrganismi e loro componenti (es. enzimi) in grado di funzionare in modo ottimale in queste condizioni. Analogamente si potrebbero isolare micorganismi ed enzimi in grado di bonificare siti contaminati da inquinanti di diversa natura.

Acknowledgements

Progetto MBLAB

Università di Bari
Ernesto Picardi
Ilenia Boria

IBBE-CNR
Francesca De Leo

ITB-CNR
Monica Santamaria
Marinella Marsano
Sabino Liuni
Domenica D'Elia
Giorgio Grillo
Flavio Licciulli

Università di Milano
David Horner
Carmela Gissi
Flavio Mignone