

Population Genomics

António Rodrigues (PDBC 2008)

Bruno Santos (PDBC 2008)

- 1 Motivation and Introduction
- 2 1000 genome project
- 3 New generation sequencing methods
- 4 Assembly: Overview and Example
- 5 Applications:
Sequencing of single individuals
Population Genomics and Geography

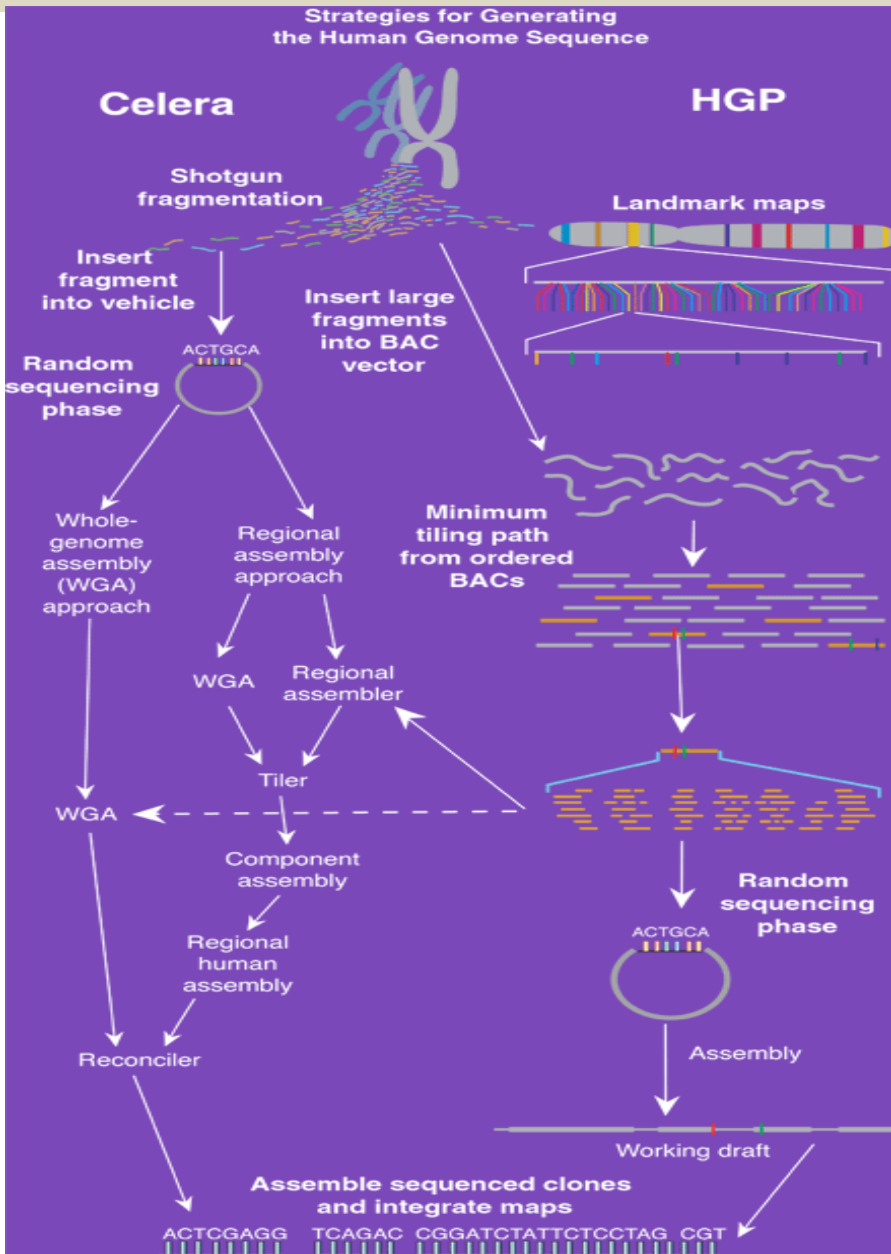
- The sequencing of the reference human genome was the capstone for many years;
- Understanding genetic diversity may reveal new insights into disease and drug response;
- Genomic size should be positive correlated to the complexity of the organism;
- The Human Genome contains about 6 gigabases;
- Human genome is 200 times larger than yeast but 200 times smaller than *Amoeba dubia*;
- This C-Paradox is now solved;
- Only less than 5% of human genome is coding sequence, repeats constitute more than 50% of the genome;

- International research consortium will sequence the genomes of at least 1000 people from around the world;
- Create the most detailed and medically useful pictures of human genome variation;
- Any two humans are more than 99% equal at genetic level;
- Variation may explain individual differences in susceptibility to diseases, responses to drugs or reaction to environmental factors;
- The HapMap project and related has already discovered more than 100 regions of the genome containing genetic variations associated with common human diseases.

- Produce a catalog of variants present at 1% or greater frequency in the human population;
- Down to 0.5 percent or lower within genes;
- Increase sensitivity of disease discovery by 5 fold across the genome and 10 fold within gene regions;
- Provide better understanding of very rare genetic diseases (<1 in 1.000 people);
- Understand contribution of common variants to most common diseases like diabetes and heart diseases;
- Identify SNP but also large differences like rearrangements, deletions or duplications

- First Phase (3 Pilot studies expected to take 2 years):
 - Sequencing the genomes of two nuclear families (both parents and adult child) at average deep coverage of 20;
 - Sequencing of 180 people at low coverage that averages 2 passes of each genome;
 - Sequencing exons of about 1.000 genes in 1.000 people;
 - Deliver 8.2 billion bases per day, more than 2 human per 24h;
- At full speed this project will generate more sequences in two days than at was added to public databases for all the past year;
- In total will generate 6 trillion DNA bases, 60 fold more sequence data than has ever been deposited in public databases;

- The first thousand samples will come from those used for the HapMap and will need to be extended;
- No medical or personal information will be collected;
- Only the population from the sample came from is known;
- Among the populations are:
 - Yoruba in Ibadan, Nigeria
 - Japanese in Tokyo;
 - Chinese in Beijing;
 - Utah residents with ancestry from northern and western Europe;
 - Luhya in Webuye, Kenya;
 - Maasai in Kinyawa, Kenya;
 - Tuscany, Italia;
 - Guajari Indians in Houston;
 - Chinese in metropolitan Denver
 - People of Mexican ancestry in Los Angeles;
 - People of African ancestry in the southwestern United States;



- Sequencing of the reference human genome took many years and was done using BAC clones;
- Produced a single contiguous stretch of high quality sequence (<1error per 40.000 bases)
- Since then sequencing have moved to WGS;
- The primary data production has relied in the same type of capillary sequencing instruments as for HGP;

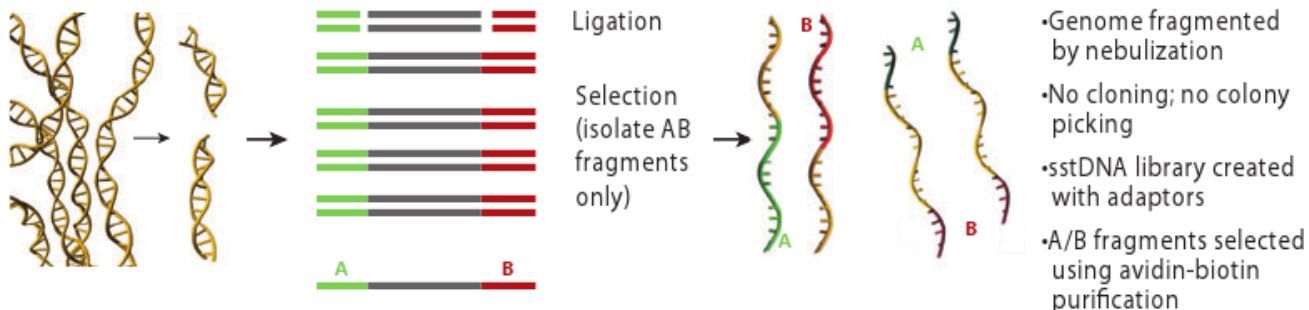
- Invention and commercial introduction of several revolutionary approaches to genome sequencing;
- 100-next generation sequencing-related manuscripts have appear;
- Improved genome sequencing timing and costs;
- Three platforms for massively parallel DNA sequencing read production are in reasonable widespread:
 - Roche 454/FLX sequencer;
 - Illumina/Solexa Genome Analyzer;
 - Applied Biosystems SOLiD;
- Two new are announced:
 - Helicos Heliscope;
 - Pacific Biosciences SMRT.

Roche/454 FLX Pyrosequencer

a

DNA library preparation

4.5 hours

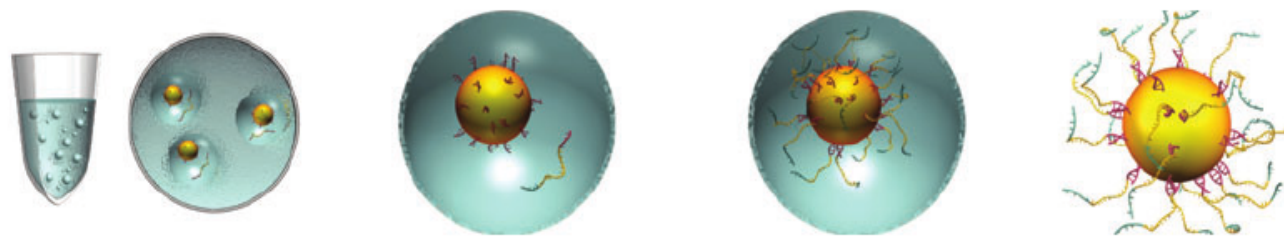


gDNA → sstDNA library

b

Emulsion PCR

8 hours



Anneal sstDNA to an excess of DNA capture beads

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

Clonal amplification occurs inside microreactors

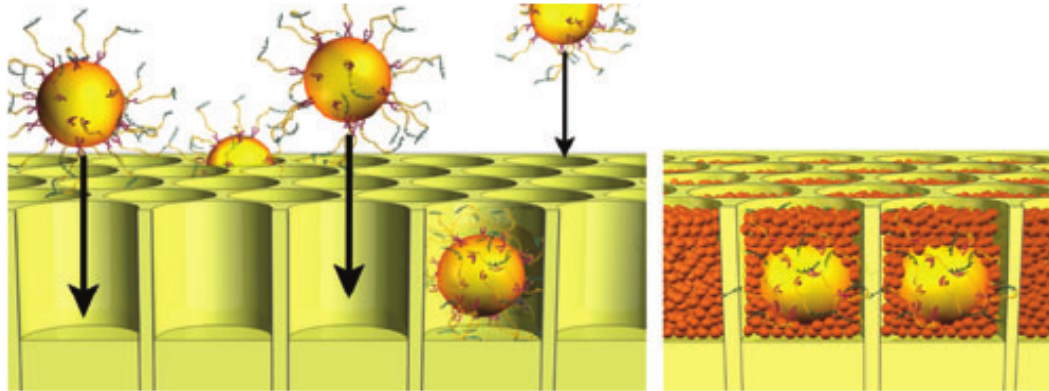
Break microreactors and enrich for DNA-positive beads

sstDNA library → Bead-amplified sstDNA library

Roche/454 FLX Pyrosequencer

Sequencing

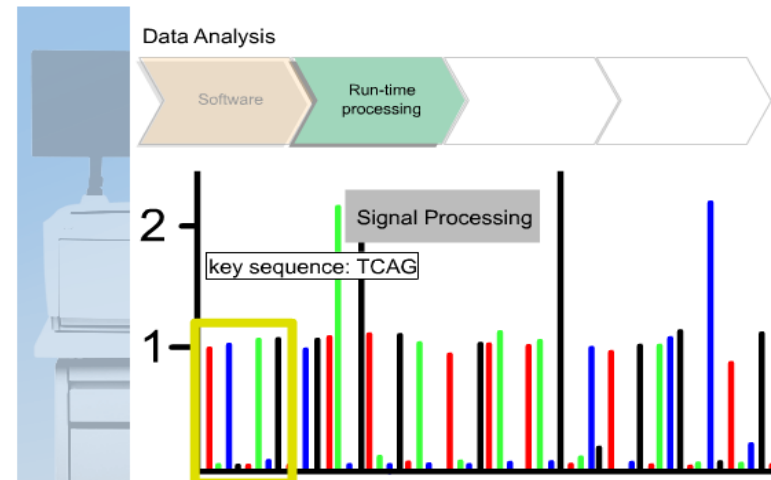
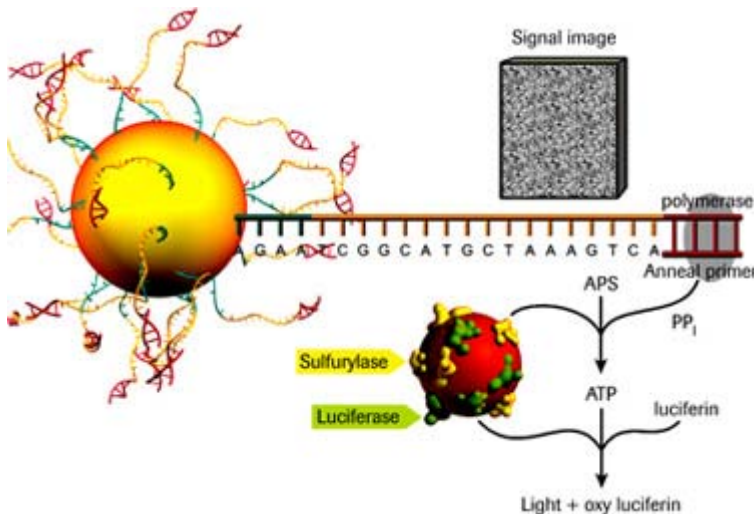
7.5 hours



- Well diameter: average of 44 μm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

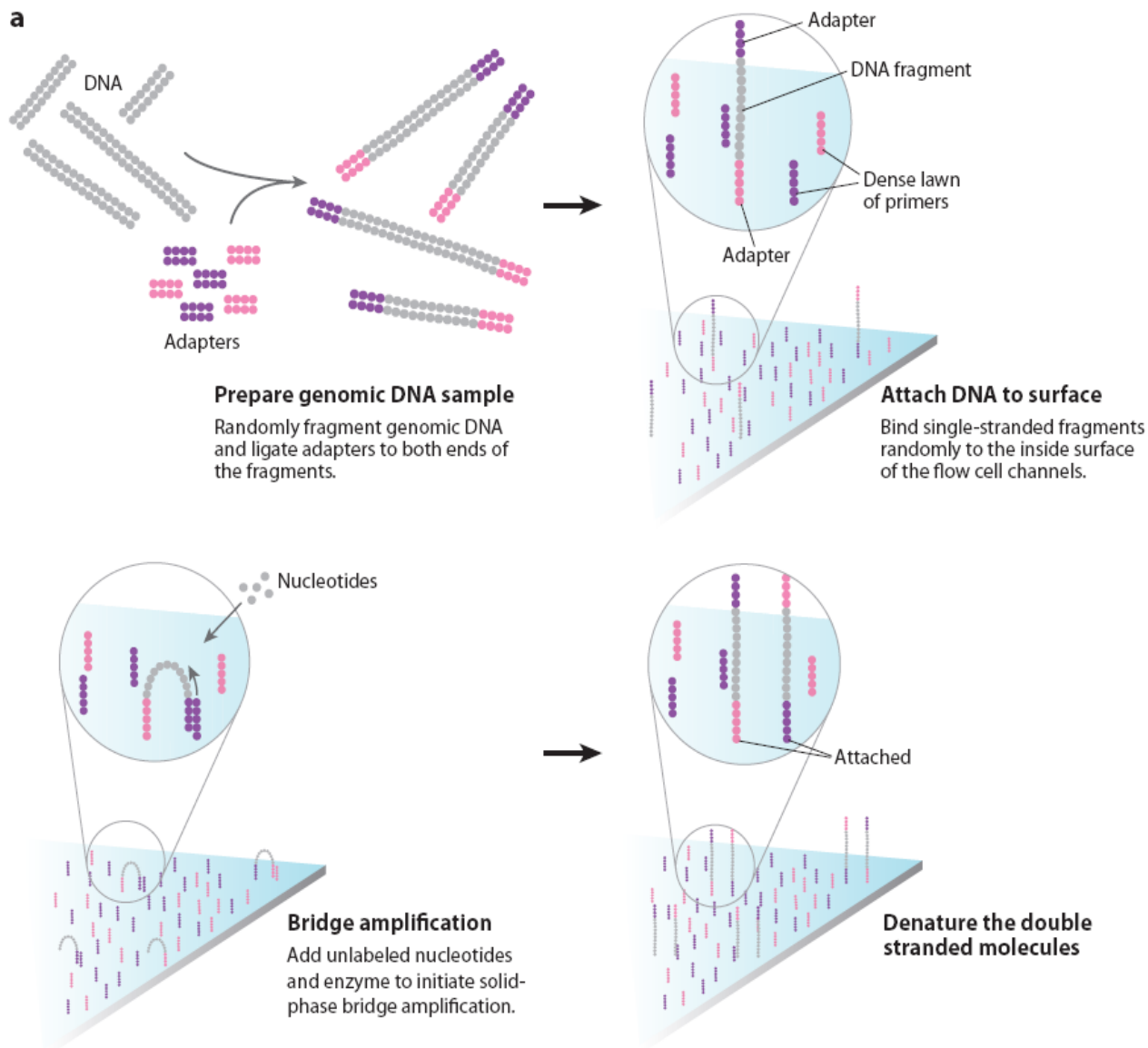
Amplified sstDNA library beads

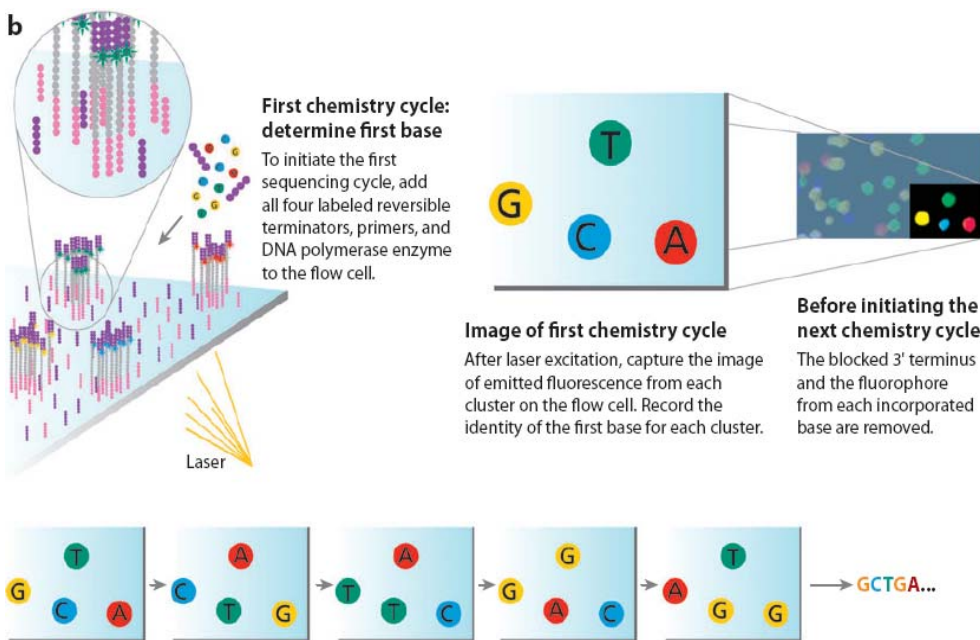
Quality filtered bases



- Cannot properly interpret long stretches (>6) of the same nucleotide;
- Prone to base insertion and deletion during base calling;
- Substitution errors are rarely encountered;
- 400-600 million high-quality, filter-passed bases per run;
- Average length of reads = 400 bases;
- 1 million high-quality reads per run;

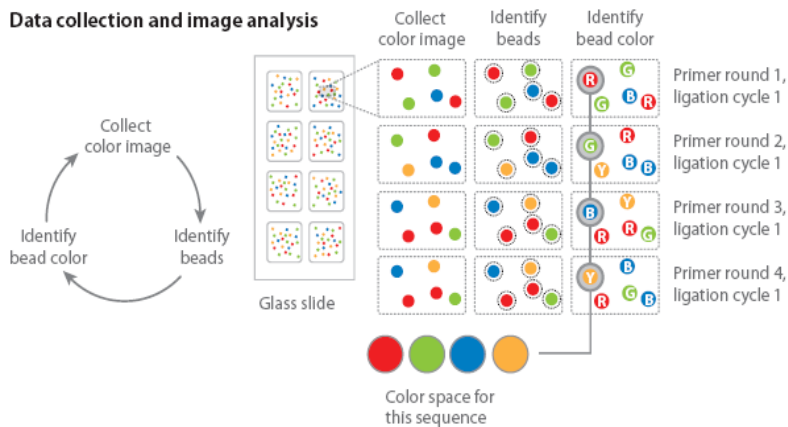
Illumina Genome Analyzer



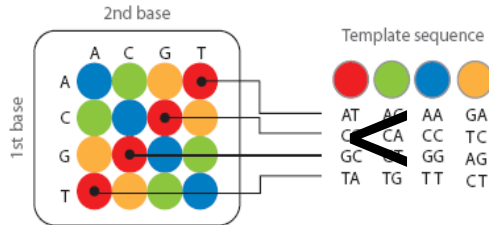


- Sequencing by synthesis solves problem with homopolymers;
- Permits discrete read lengths of 25-35bp;
- 1~3 Gb of data per run;
- 2.5 Gb of high quality data;

b Data collection and image analysis

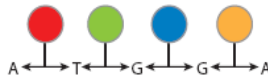


Possible dinucleotides encoded by each color

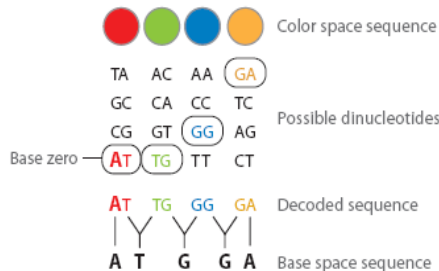


Double interrogation

With 2 base encoding each base is defined twice



Decoding



- Greater than 99,94% accuracy due to the 2 base encoding;
- Over 20 gigabases per run;
- Can sequence an human genome for \$10,000;
- Read length can now go to 50 bp or 2x50bp.

- Richness of repeats in Eukaryotes poses great challenges for fragment assembly
- These repeats can cause misassemblies specially when using whole genome shotgun (WGS) methods
- Repeats can:
 - Be missed and left as gaps
 - May be collapsed
 - Cause misjoin of nonadjacent fragments

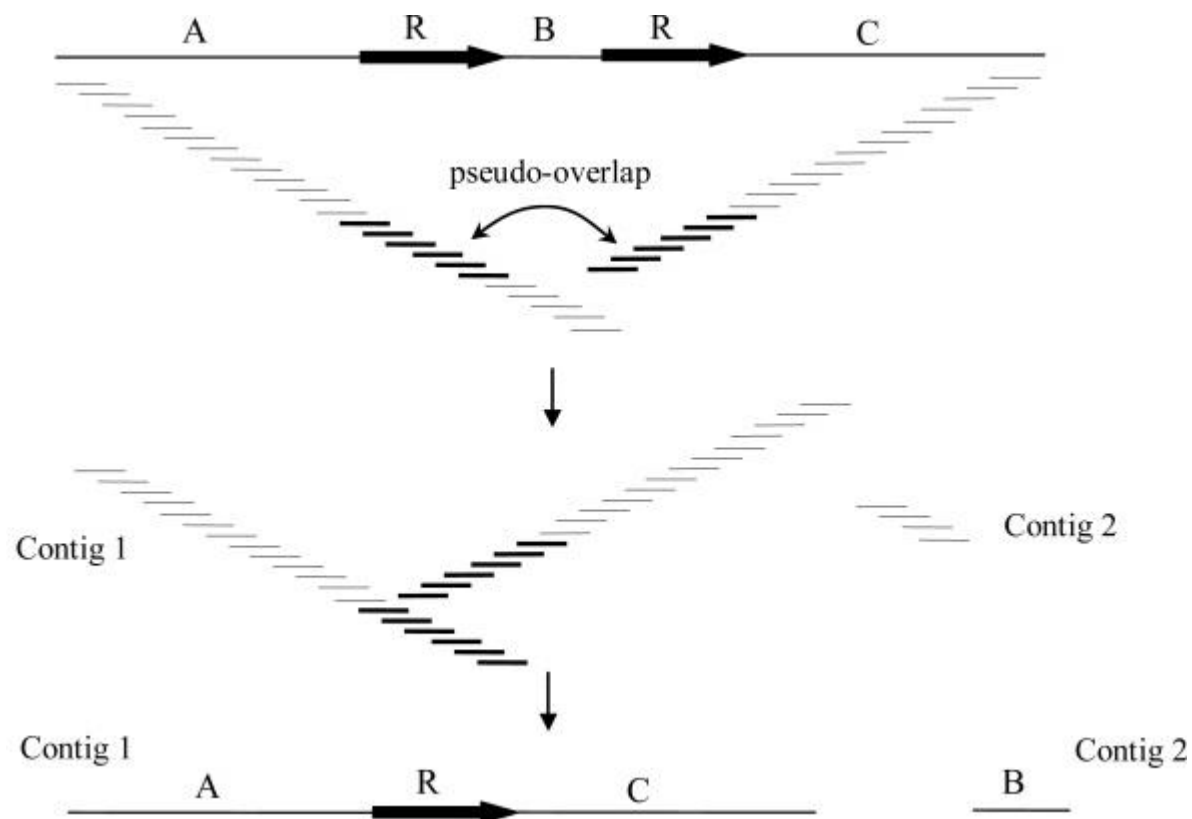
- To sequence long DNA molecules a shot gun strategy is adopted
- This involves breaking the target DNA sequence in overlapping fragments, we obtain short pieces of DNA called **reads**
- Based on the overlap regions the reads must be put together in order to reconstruct the original genome, this is done by an automated computer program called **assembler**

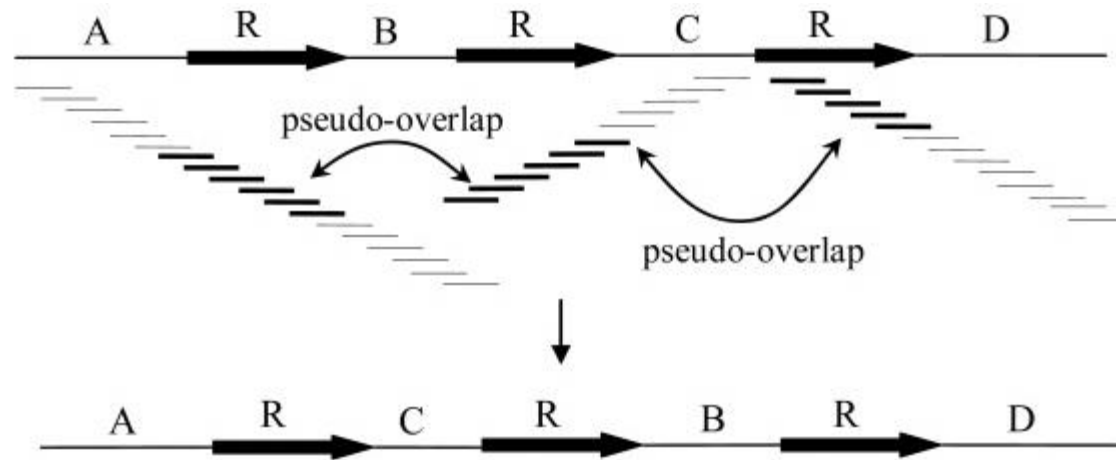
- **'Overlap-layout consensus'** paradigm
- Two steps: **overlap step** & **layout step**, it might also involve a **scaffolding step**
- **Contigs**: overlapping reads
Read coverage :(ratio between the length of the reads and the length of the genome)

- Different copies of repeats are very similar this can originate **pseudo-overlaps**
- Pseudo-overlaps may cause:
 - **base-calling errors:**
Repeats are mistakenly placed
 - **false rearrangements**
Large-scale rearrangements of DNA segments

Assembly problems |

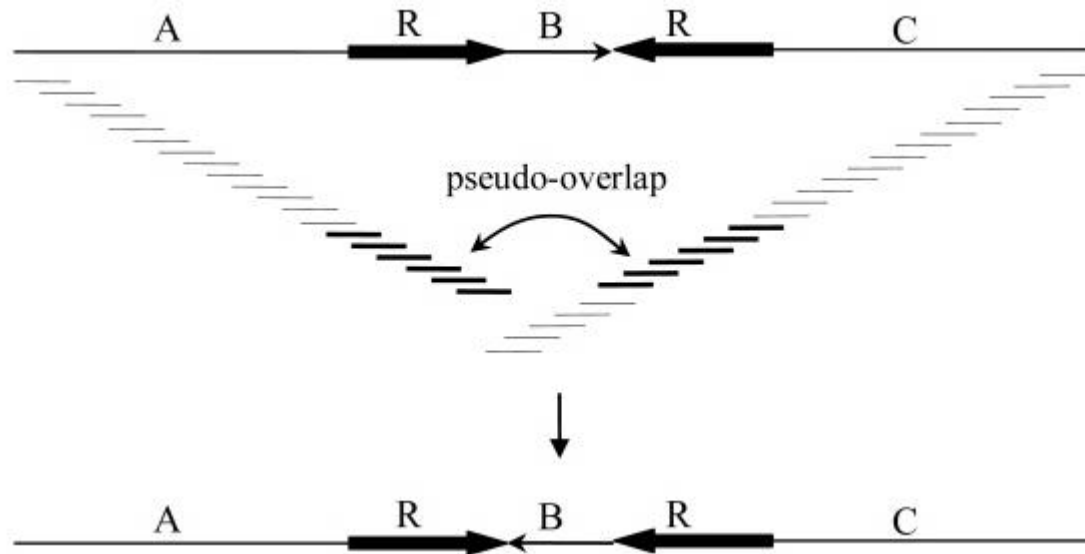
- Repeats in the genomes poses problems for the assembly of fragments
- Two repeat copies can be collapsed





- Three repeats can cause a misassembling of the inner segments

Assembly problems |

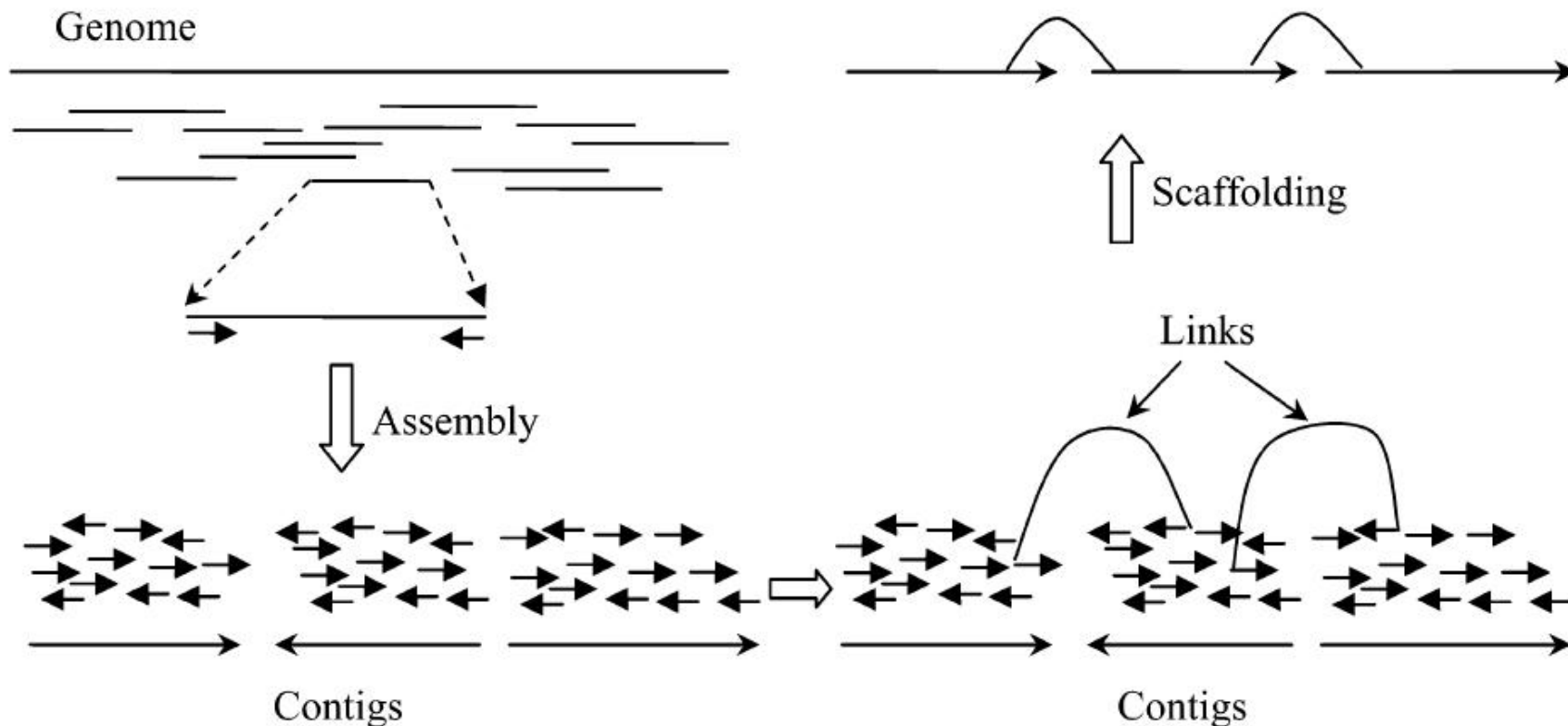


- Two inverted repeat copies can cause a misassembled sequence

Repeat masking |

- Repeat masking: **detect** and **not to assemble** repeats
- Remaining reads from unique regions are assemble into **contigs (unitigs)**
- Repeats databases can be used to detect and mask **known repeats**, otherwise statistical methods can be used
- These contigs are grouped into **scaffolds** with order and orientation (some assemblers contain this scaffolding step).

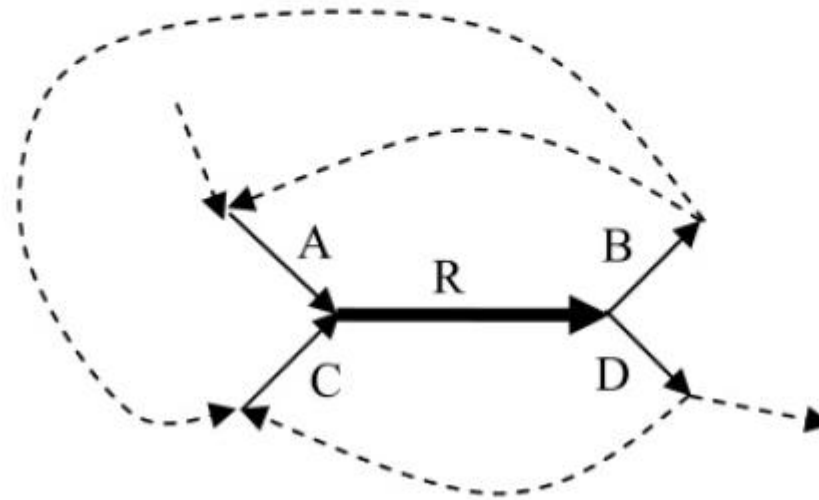
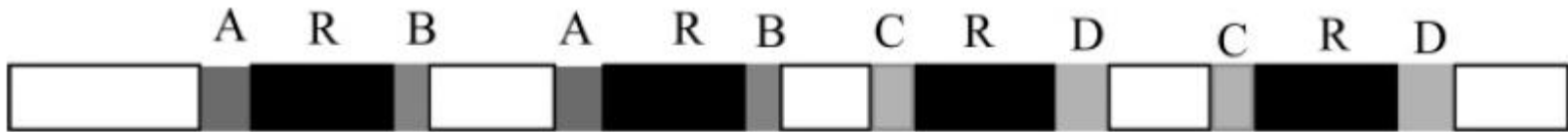
Scaffolding step |



- This process involves a scaffolding step

- Not all repeats can be detected and masked
- Different approaches are needed: **Eulerian path approach**
- **Eulerian path approach**: represent repeats using a **repeat graph**
- **Eulerian path approach**: a path that visits each edge in the graph once and only once

Repeat graph |

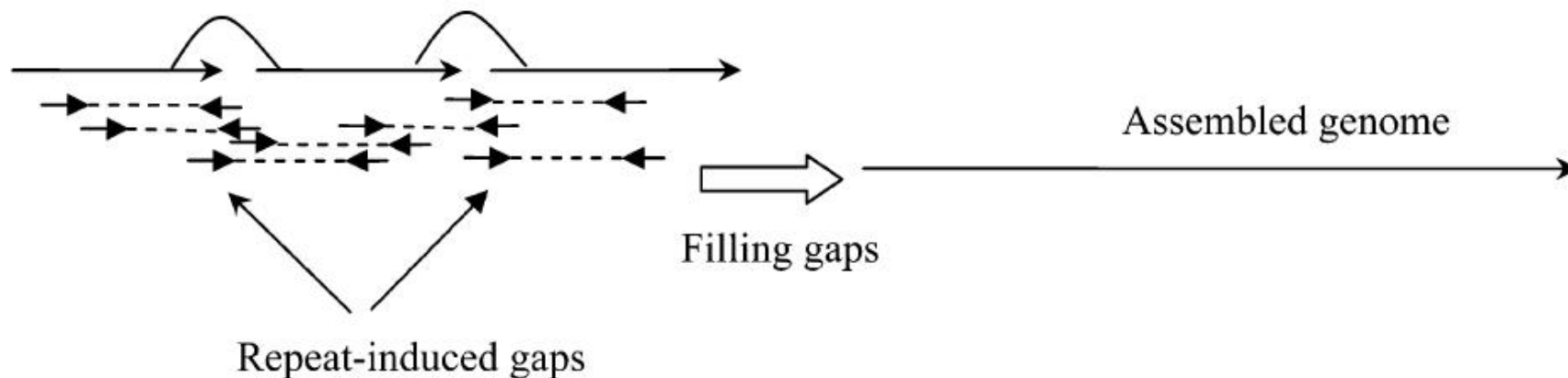


- Repeat graph example

Repeat resolution with Doubled-Barreled Data |

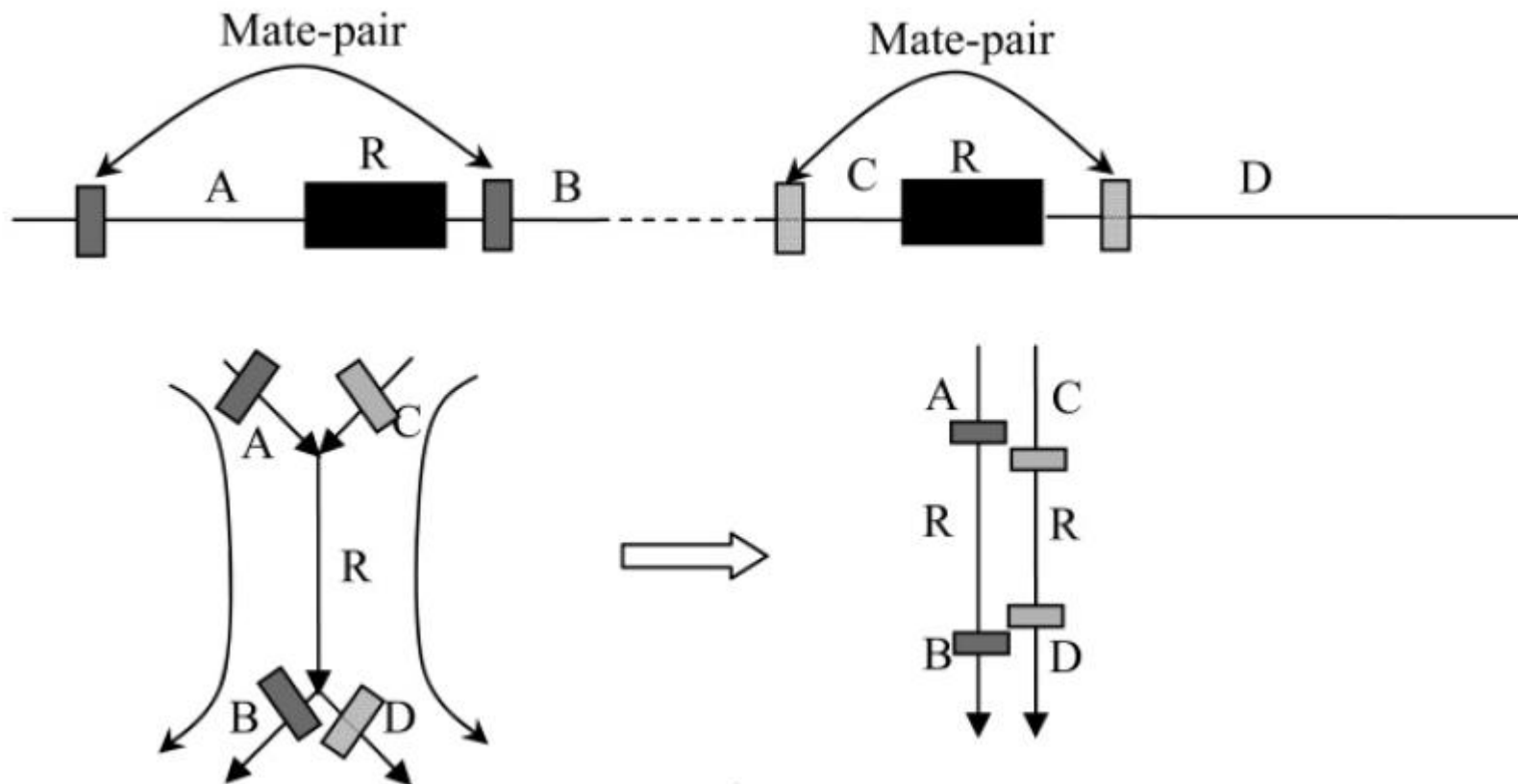
- The repeat graph can be highly complex
- The shotgun strategy can be improved using **double-barreled sequencing**
- **double-barreled sequencing**: it is done by obtaining pairs of reads (mate-pairs) separated by a medium-insert clone

Overlap-consensus layout |



- Masked read can fill the gaps of the scaffold. This is done using the mate pairs

Repeat graph |



- Within the repeat graph framework, double-barred data can be used to eliminate some repeat edges

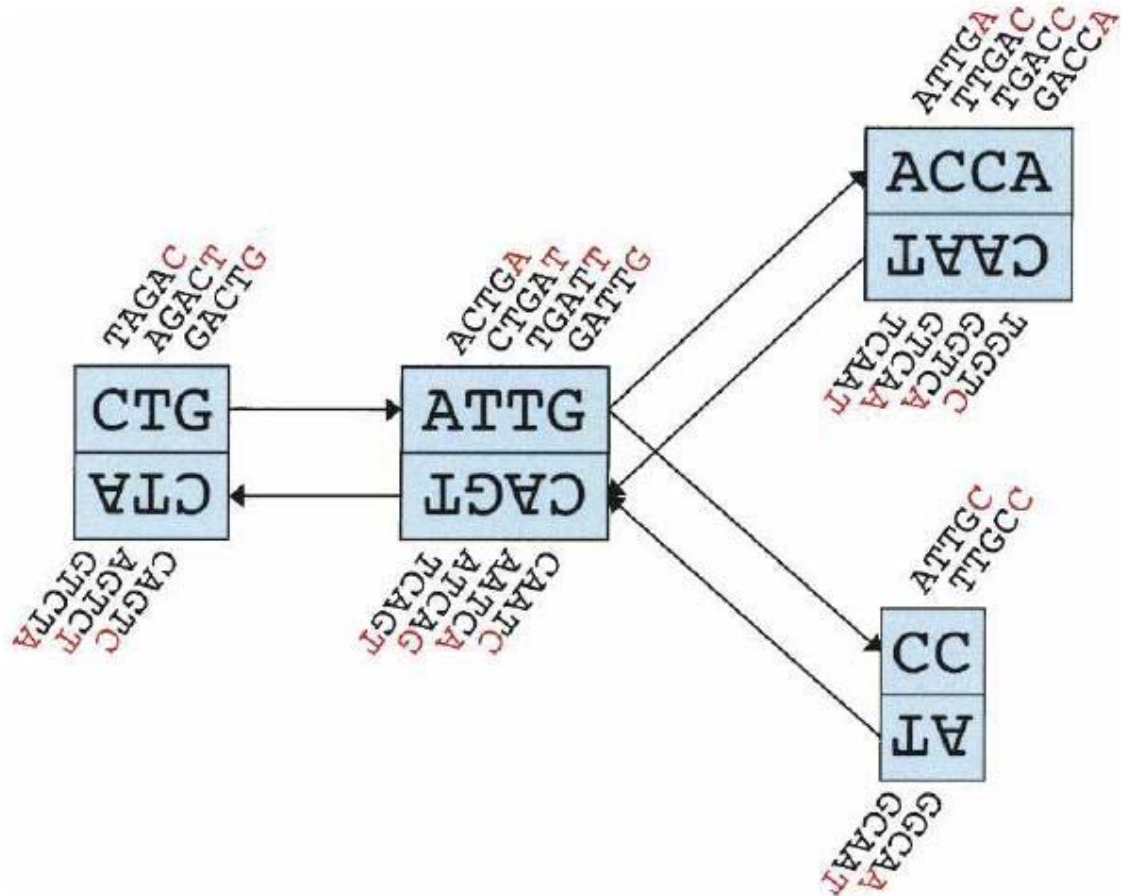
- Velvet: algorithms for de novo short read assembly using de Bruijn graphs
- Each read is represented as a node and each overlap between appropriate nodes
- **Euler assembler** adopts a different approach and uses Bruijn graph.
- **Bruijn graph**: elements are organized around k-mers, words of k nucleotides. Reads are mapped as paths through the graph.

These algorithms manipulates Bruijn graphs to both eliminate errors and resolve repeats

Bruijn graph |

- Each node N represents a series of overlapping k -mers
- Adjacent k -mers overlap by $k-1$ nucleotides
- Marginal information of a k -mer is its last nucleotide
- The final nucleotides are the sequence of the node $s(N)$
- Each node N has an attached twin node \tilde{N} which has reverse complement k -mers to handle opposite strands
- Nodes are connected by a directed 'arc'

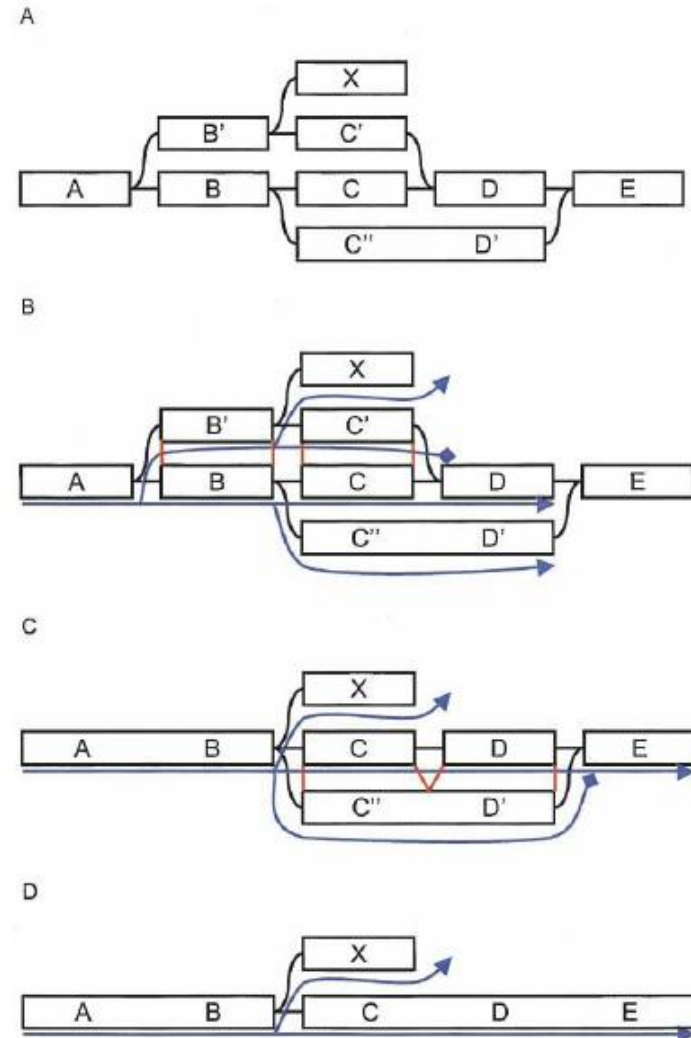
- The last k-mer of an acrs's origin node overlaps with the first of its destination node
- The blocks are symmetric therefore an arc from A to B implies a symmetric arc from $\sim B$ to \tilde{A}



- Hash table that associates k-mers with read
- This representation is called 'roadmap'
- Erroneous data create three type of structures:
 - Tips
 - **Bubbles**
 - Erroneous connections

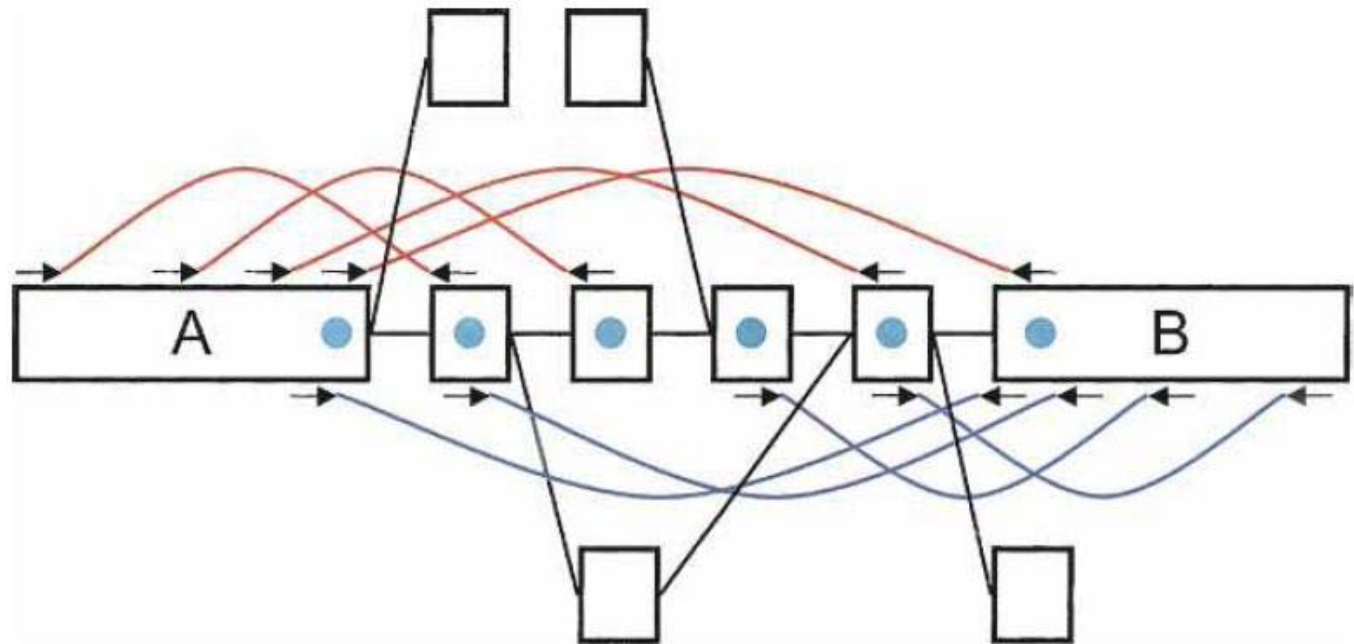
Error removal – Tour Bus error correction |

- ‘Bubbles’ can be solved using an algorithm called ‘Tour bus’
- It uses a Dijkstra-like breadth-first search
- Sequences are extracted from the paths and are aligned. If they are similar they are merged



Assembly: Algorithm

Repeats short read pairs – Breadcrumb algorithm |



- Long contigs are joined by several paired reads
- Nodes that are paired both to A or B are marked

- Levy, S. *et al.* **The Diploid Genome Sequence of an individual Human.** PloS Biology (2007)
- Wheeler, D. A. *et al.* **The complete genome of an individual by massively parallel DNA sequencing.** Nature (2008)
- Wang, J. *et al.* **The diploid genome sequence of an Asian individual.** Nature (2008)

- Produced from ~32 million random DNA fragments;
- Sequenced by Sanger technology;
- Assembled into 4,528 scaffolds, comprising 2,810 bases of contiguous sequences;
- Approximately 7,5-fold coverage;
- Projected developed over a 10-year period.

- The two current versions of human genome are a composited derived from haploids of numerous donors;
- Both versions almost exclusively report DNA variation in the form of SNP;
- Smaller-scale (<100 bp) insertion/deletion sequences
- Large-scale structural variants also contribute to human biology and disease;
- The initial draft of genomes provide an excess of 2.4 million SNPs;
- With current SNP-based genome wide association studies rely on population data and therefore can be uninformative or misleading

- J. Craig Venter
- Born on 14 October 1946;
- Caucasian male;
- Personal, medical and prototypic traits data were collected;
 - Has 2 brothers, 1 sister and one biological son;
 - His father died at age 59 of sudden cardiac arrest;
 - There are documented cases of chronic disease in family such as hypertension and ovarian and skin cancer;
 - Genealogical records can be traced back to 1821 (paternal) and the 1700s (maternal in England)
- No obvious chromosomal abnormalities.

- Assembled with a modified version of celera assembler;
- Improving coverage and improving assembler resulted in 68% decrease in the number of gaps within scaffolds;
- Resulted in 4,528 scaffolds;
- Genomic variation was observed by two approaches:
 - Heterozygous alleles within diploid genome;
 - Comparison between HuRef and NCBI version 36 human reference assembly;

Table 2. Summary of HuRef Assembly Statistics and Comparison to the Human NCBI Genome

| Assembly | Assembly Subset | Number of Scaffolds | Number of Contigs | Gaps within Scaffolds | ACGT Bases | Span |
|------------------|-------------------------|---------------------|-------------------|-----------------------|---------------|---------------|
| NCBI Chromosomes | N/A | 279 | N/A | N/A | 2,858,012,806 | 3,080,419,480 |
| NCBI All | N/A | 367 | N/A | N/A | 2,870,607,502 | 3,093,104,542 |
| WGSA Chromosomes | N/A | 4,940 | 211,493 | 206,553 | 2,659,468,408 | 2,993,154,503 |
| HuRef Assembly | Chromosomes | 1,408 | 66,762 | 66,354 | 2,782,357,138 | 2,809,547,336 |
| | Scaffolds \geq 100 kb | 553 | 65,932 | 65,379 | 2,779,929,229 | 2,806,091,853 |
| | Scaffolds \geq 3 kb | 4,528 | 71,343 | 66,815 | 2,809,774,459 | 2,844,046,670 |
| | All scaffolds | 188,394 | 255,300 | 66,906 | 3,002,932,476 | 3,037,726,076 |

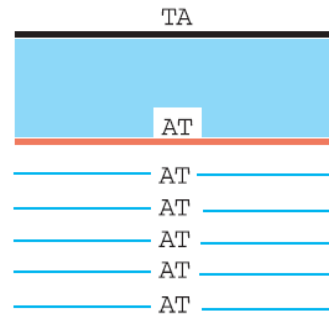
Genomic variation

homozygous SNP



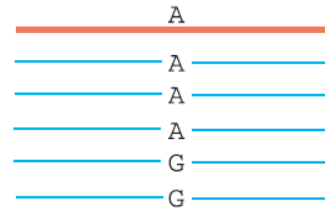
variant: A/A

homozygous MNP



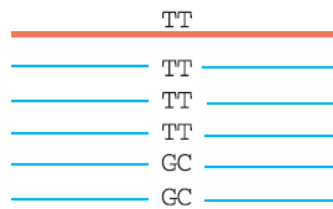
variant: AT/AT

heterozygous SNP



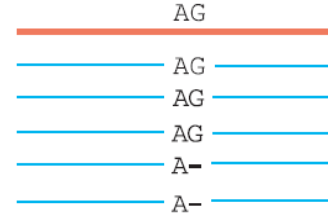
variant: A/G

heterozygous MNP



variant: TT/GC

heterozygous indel



variant: -/G

complex



variant: TT/C-

homozygous insertion



variant: —

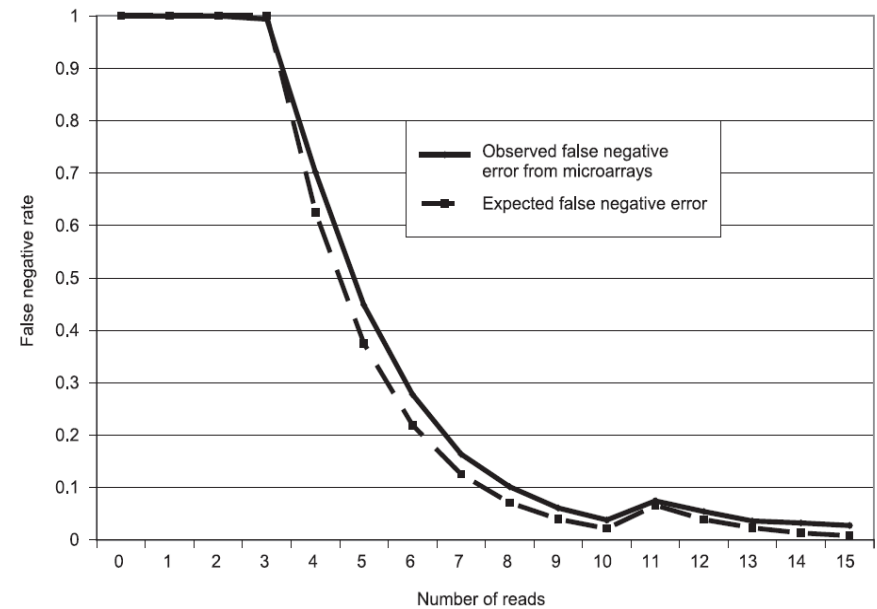
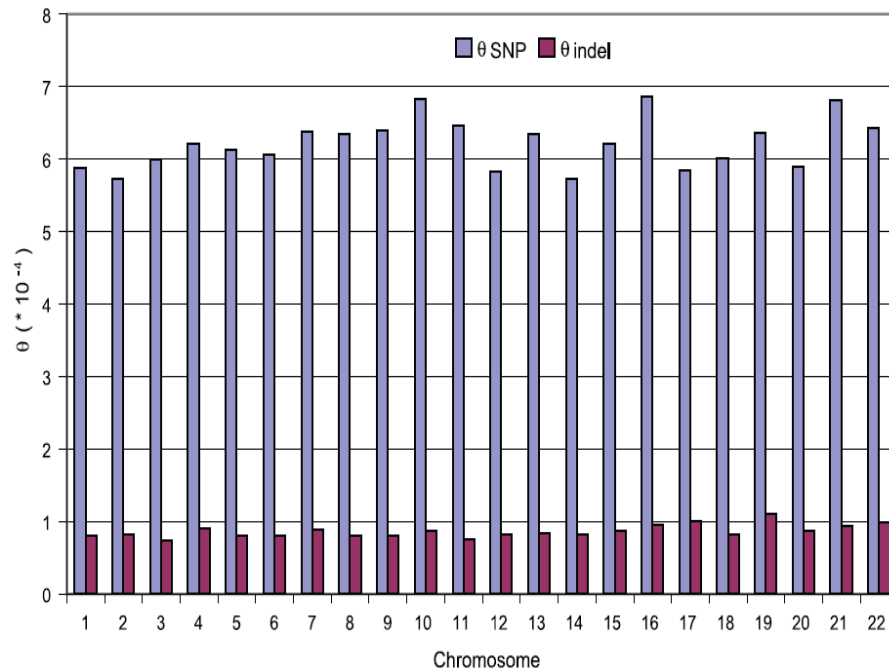
homozygous deletion



variant: —

Table 6. Summary of Variant Types Identified in the HuRef Genome Assembly

| Type | Number of Variants | bp Length | Min | Max | Mean | % Variants in Repeat Sequence |
|----------------------|--------------------|------------|-----|---------|--------|-------------------------------|
| heterozygous SNP | 1,762,541 | 1,762,541 | 1 | 1 | 1 | 52 |
| homozygous SNP | 1,450,860 | 1,450,860 | 1 | 1 | 1 | 56 |
| heterozygous MNP | 38,985 | 227,531 | 2 | 206 | 5.8 | 52 |
| homozygous MNP | 14,838 | 31,590 | 2 | 22 | 2.1 | 69 |
| heterozygous indel | 263,923 | 635,314 | 1 | 321 | 2.4 | 71 |
| Complex | 28,179 | 330,803 | 2 | 571 | 11.7 | 70 |
| homozygous insertion | 275,512 | 3,117,039 | 1 | 82,711 | 11.3 | 74 |
| homozygous deletion | 283,961 | 2,820,823 | 1 | 18,484 | 9.9 | 78 |
| inversion | 90 | 1,914,477 | 7 | 670,345 | 21,272 | 98 |
| Total | 4,118,889 | 12,290,978 | | | | |



- 17% of the coding genes encode differential proteins.
- 44% of genes at least 1 heterozygous variant in the UTR or coding region;
- Almost half of the genes could have differential states;
- Donor is heterozygous in the polymorphic trinucleotide repeat located in the Huntington disease;
- The donor is heterozygous for variants in alleles associated with cardiovascular diseases that present a lower risk of this disease;
- Have also been found novel changes for which biological implications are unknown;
- Inconsistencies between detected genotypes were also found, should have lactose tolerance.

- 44% of the annotated genes have at least 1, or often more, alterations within them;
- 78% of the variants detected are SNPs;
- The remaining 22% non-SNPs account for 74% base variants;
- Copy number variation also shown variation within the genome;
- A minimum of 0,5% variation exists between two haploid genomes;
- The repeat regions were ignored so a very large may escaped from the analysis;
- Further family sampling would be required to determine the relevance between genotype and phenotype.

- Published 6 months after;
- Sequenced using 454 FLX sequencer;
- 24,5 billion bases created, resulting in 93,2 million reads aligned to the reference genome sequence;
- Reference genome was then covered to an average of 7.4 fold coverage;
- Cost less than US\$1 million compared to the US\$100 spent with Craig Venter;
- Completed in two months;

Table 1 | Single nucleotide variation in 454 reads

| Subject | Filter* | Total variation | Known† | Novel |
|---------|---------|-----------------|-----------|------------|
| Watson | Raw | 14,829,087 | 3,283,273 | 11,545,814 |
| | 1 | 4,427,488 | 2,815,322 | 1,612,166 |
| | 2 | 3,971,513 | 2,752,991 | 1,218,522 |
| | 3 | 3,322,093 | 2,715,296 | 606,797 |
| Venter‡ | 4 | 3,470,669 | 2,822,902 | 647,767 |

* Filters: raw, all base substitution from cross_match alignments; 1, $S_v > 28$ (see Methods); 2, filter 1 plus ratio of variant to total coverage > 0.2 ; 3, filter 2 plus eliminate SNPs close to homopolymer runs > 5 bp; 4, (Venter) Phred (ref. 20) $Q > 15$, ratio of variant to total coverage > 0.2 .

† Variants found in build 126 of dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>).

‡ SNPs found in genome of Venter: see ref. 2 and supplementary material therein.

Table 2 | Microarray validation of 454 SNPs

| Affymetrix genotype* | Affymetrix SNP array | 454 Sequence† | Agreement (%) |
|----------------------|----------------------|---------------|---------------|
| Homo ref. | 254,753 | 253,348 | 99.4 |
| Homo var. | 104,547 | 99,387 | 95.1 |
| Hetero | 135,413 | 102,702 | 75.8 |

* Homo ref., homozygous for reference allele; homo var., homozygous for the variant allele; hetero, heterozygous.

† The genotype based on the alleles observed in 454 reads at each position of an Affymetrix marker.

- Sequenced using Solexa/Illumina;
- Paired-end libraries were used;
- Read length average of 35base pairs;
- 3.3billion reads were collected;
- 117.7 Gigabases;
- Aligned to the NCBI genome with SOAP (87.4% of data);
- 36x fold coverage.

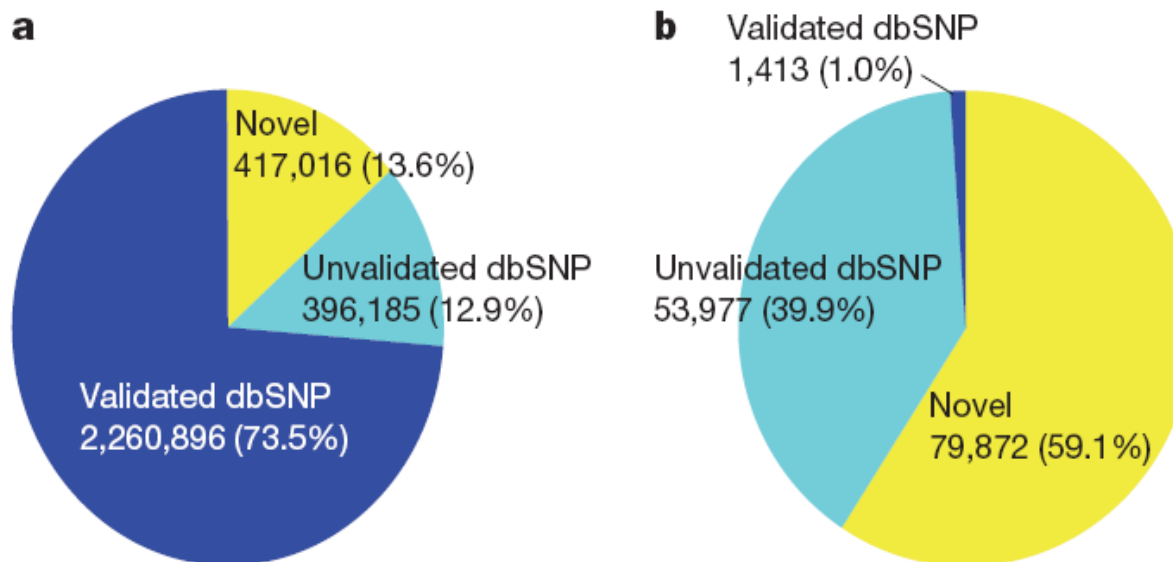


Figure 1 | The percentage of detected SNPs (a) and small indels (b) that overlap with SNPs and small indels in the dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP/>, build 128). The dbSNP alleles were separated into validated and non-validated SNPs, and the detected SNPs that were not present in dbSNP were classified as novel.

Table 1 | Data production and alignment results for the YH genome

| Data type | Number of reads | Number of mapped reads | Total bases (Gb) | Mapped bases (Gb) | Effective depth (fold) | Percentage with unique placement | Rate of nucleotide mismatches (%) |
|-----------|-----------------|------------------------|------------------|-------------------|------------------------|----------------------------------|-----------------------------------|
| SE | 2,019,025,890 | 1,921,271,902 | 72 | 64.4 | 22.5 | 83.60 | 1.62 |
| PE | 1,315,249,404 | 1,028,695,924 | 45.7 | 38.5 | 13.5 | 90.20 | 1.16 |
| Total | 3,334,275,294 | 2,949,967,826 | 117.7 | 102.9 | 36 | 86.10 | 1.45 |

Single-end (SE) and paired-end (PE) sequencing reads were aligned onto the reference assembly in NCBI build 36.1, allowing at most two mismatches or one continuous gap with a size of 1–3 bp. Effective depth was determined through the calculation of all mapped bases divided by the length of NCBI36 (excluding Ns, 2,858,013,089 bp in length). 'Unique placement' means a read had only one best placement with the least number of mismatches and gaps. The rate of nucleotide mismatches is the percentage of mismatched nucleotides over all mapped nucleotides, including sequencing errors and real genetic variations. In total, 487 million reads (14.6%) could not be aligned to the reference genome.

a

| | Total sites | In DGV | | Overlap TEs | |
|---------------|-------------|--------|----|-------------|----|
| | | No. | % | No. | % |
| Duplication | 33 | 23 | 70 | 19 | 58 |
| Inversion | 17 | 11 | 65 | 15 | 88 |
| Deletion | 2,441 | 1,613 | 66 | 1,834 | 75 |
| Other complex | 191 | 117 | 61 | 58 | 30 |
| Total | 2,682 | 1,764 | 66 | 1,926 | 72 |

- The asian individual was estimated to share alleles at 94.12% with the Asian, 4.12% with the European and 1.76% with the African population;
- Assuming an infinite-site model of neutral mutations and equilibrium of mutation and drift;
- Chinese effective population estimated to be 5,700;
- Same analysis applied to the CV and JW gives 3,300.

- Surveyed 1,495 alleles of 116 genes described in OMIM and identified one heterozygous recessive mutation for deafness disorder;
- Complex phenotypes identified several genotypes associated with tobacco addiction and Alzheimer's disease;
- The individual is a heavy smoker;
- The donor contains 9 (56.3%) of the 16 identified Alzheimer's disease risk alleles;
- With the lack of family history information is not possible to infer family history in this disease.

Genes and European geography |

- High-throughput genotyping technologies with dense geographic samples can shed light on unanswered questions regarding human population structure
- To what extent populations within continental regions exist as discrete genetic clusters or as a genetic continuum?
- How precisely one can assign an individual to a geographic location based on genetic information alone?

Genes and European geography |

- It was used genetic information of 3 192 European individuals
- Individuals were genotyped at 500 568 loci
- The geographic location was assigned according to the individual's grandparents origin, otherwise it was used the self-reported country of birth
 - SNPs with low scores were removed
 - Individuals from outside Europe
 - Individuals with grandparents from more than one origin were also removed
 - Remove individuals with SNPs in high linkage disequilibrium

Genes and European geography |

- The analysis were focused on data from 197 146 loci in 1 387 individuals
- Principal component analysis was used to produce 2-dimensional visual summary
- The structure of this plot has a notable resemblance to the geographic map of Europe:
 - Large structures like the peninsula
 - Small structures like the French, German and Italian speaking groups of Switzerland and Ireland and UK

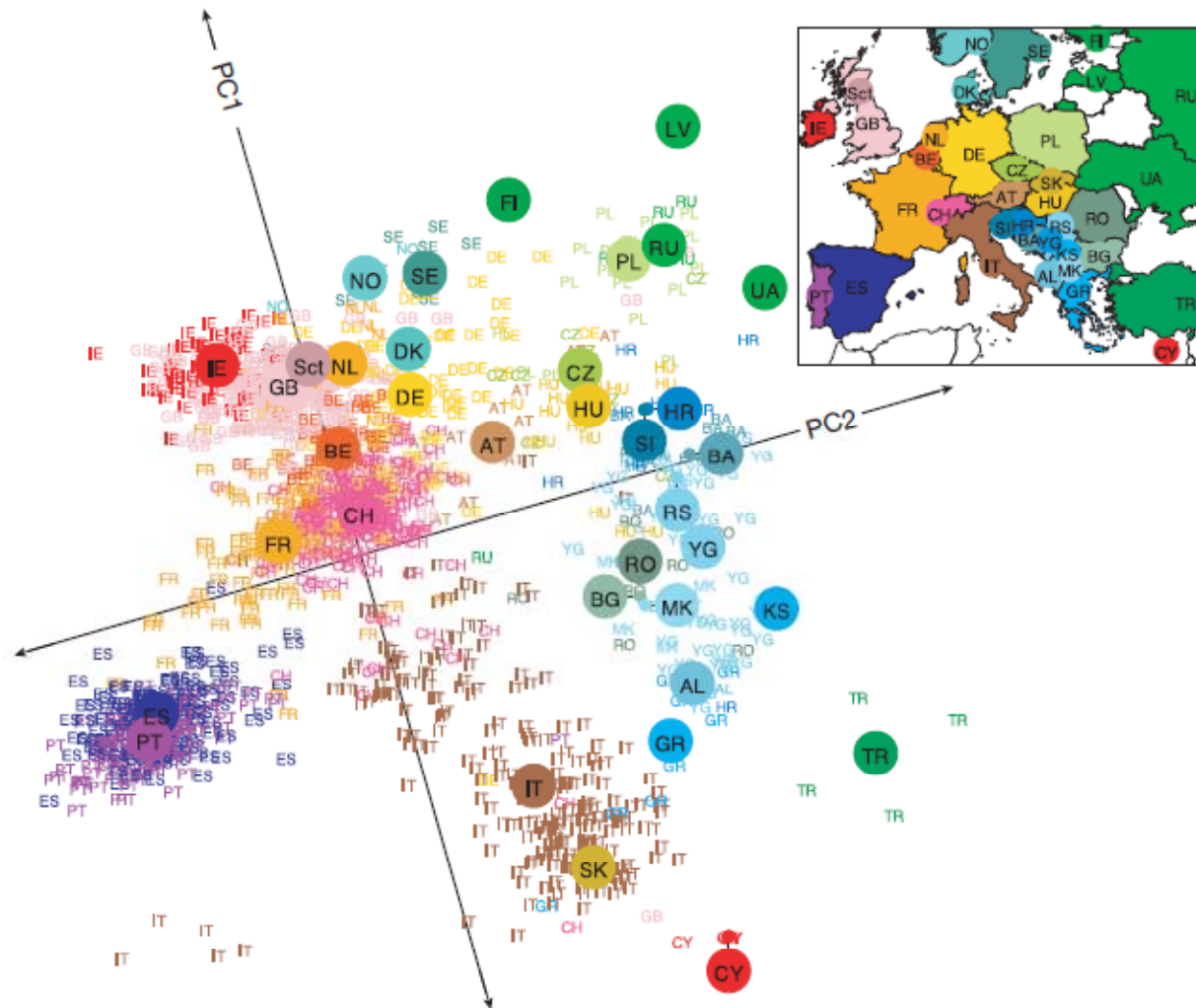
Genes and European geography |

- The results are consistent with the theoretical expectation for model where genetic similarity decays with distance (not discrete well-differentiated populations)
- PC1 is correlated with the NNW-SSE direction. This is consistent with the proposed demographic history of Europe (Genetic diversity also decreases)
- 50% of the individuals can be placed within 310 km of their reported origin
90% within 700 km

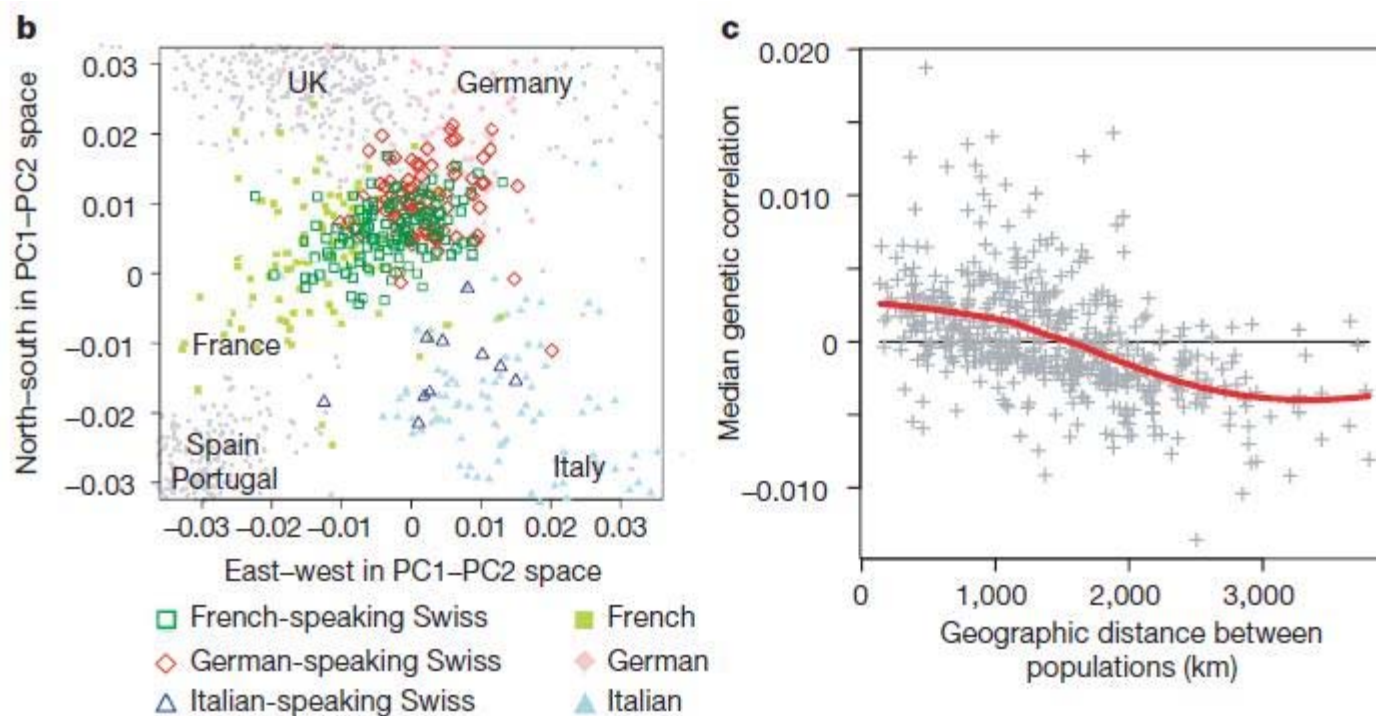
Population Genomics: Example

Genetic and geographic structure |

- Large dots represent median values for each country



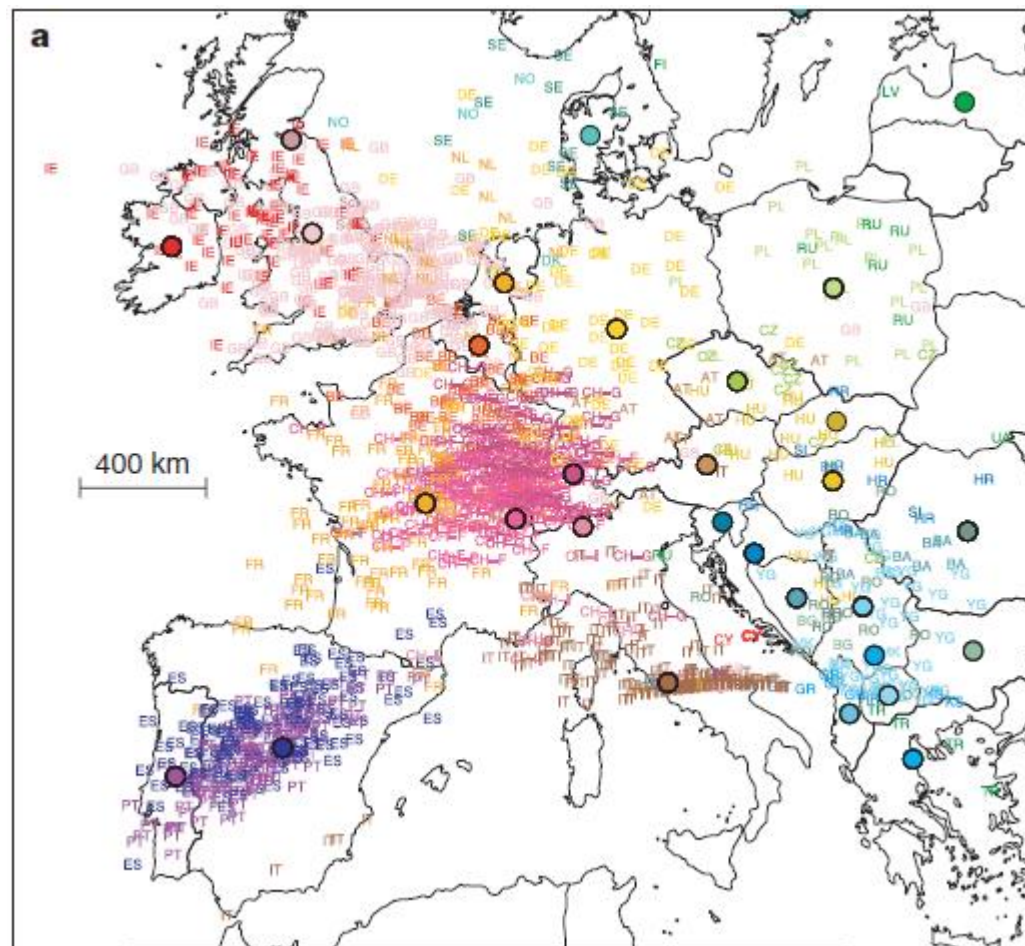
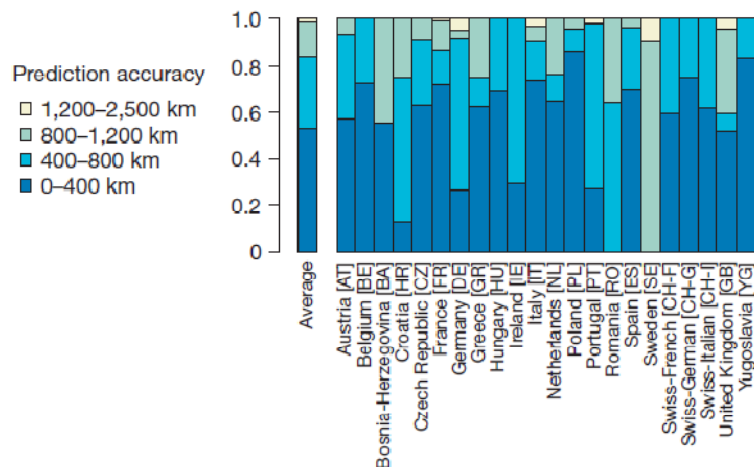
Genetic distance and geographic distance |



● Genetic distance correlated with genetic correlation

Accuracy of predictions |

- Performance decreases for populations with smaller sample sizes



END

Questions ?