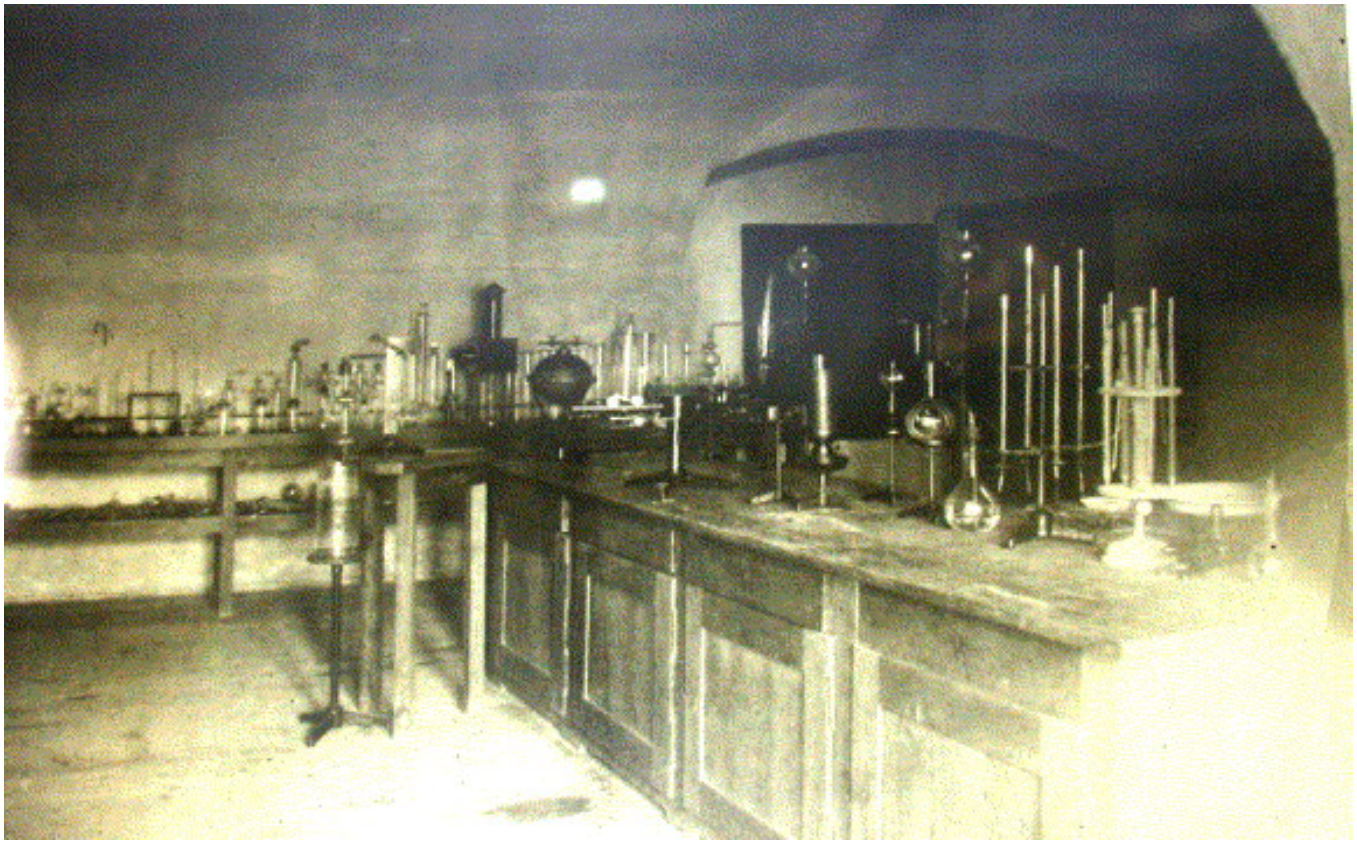
Principles of sequencing: DNA, RNA

GBIO0002

2018-11-13

Tina O’Grady

1869: discovery of DNA



Dahm 2005

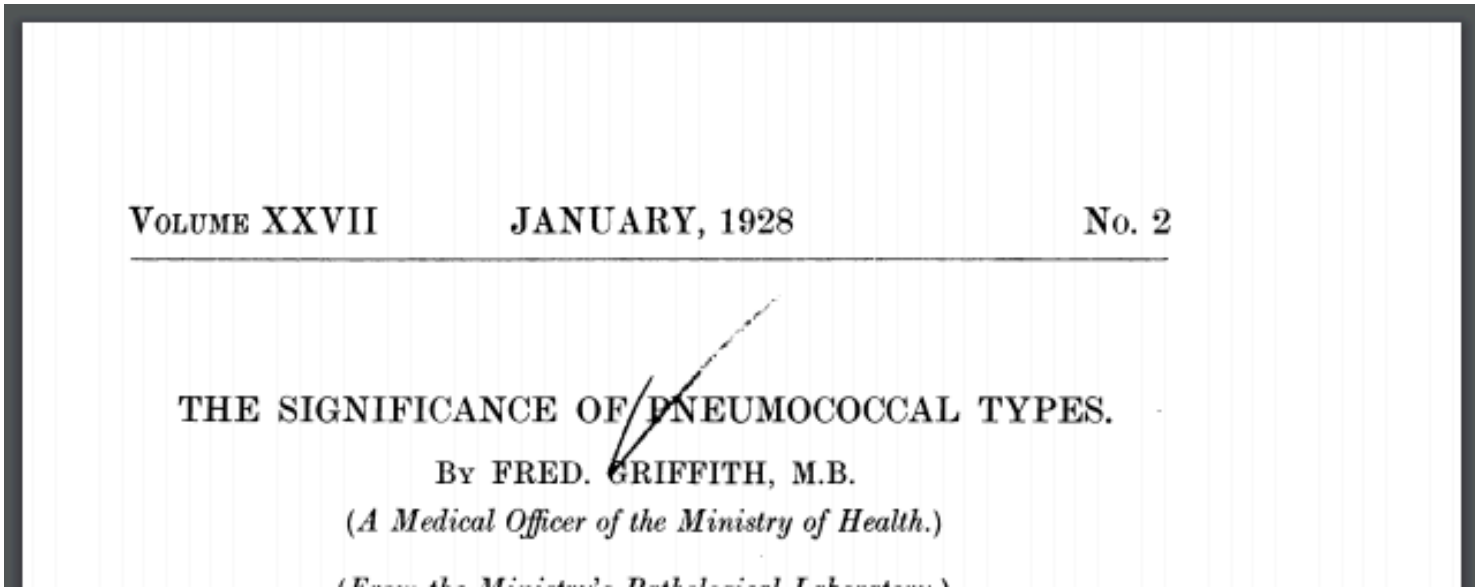
Friedrich Miescher: isolated unknown substance from leukocyte nuclei

- Contained carbon, hydrogen, oxygen, nitrogen…
- and high amounts of phosphorus, but no sulfur (so, it wasn't protein)
- Called it "nuclein"

DNA as the hereditary material

1928: Frederick Griffiths shows that dead virulent bacteria can transform living non-virulent bacteria, making it virulent

What is the hereditary material that allows this?



VOLUME XXVII JANUARY, 1928 No. 2

THE SIGNIFICANCE OF PNEUMOCOCCAL TYPES.

BY FRED. GRIFFITH, M.B.

(*A Medical Officer of the Ministry of Health.*)

DNA as the hereditary material

1944: Oswald Avery isolates many different substances from virulent bacteria and applies them to nonvirulent bacteria

"Preparation 44" transforms the nonvirulent bacteria.

It is high in phosphorus, and more tests show it is DNA.

STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES

INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III

BY OSWALD T. AVERY, M.D., COLIN M. MACLEOD, M.D., AND MACLYN MCCARTY,* M.D.

(*From the Hospital of The Rockefeller Institute for Medical Research*)

1953: the Double Helix

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. WATSON
F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems,
Cavendish Laboratory, Cambridge.

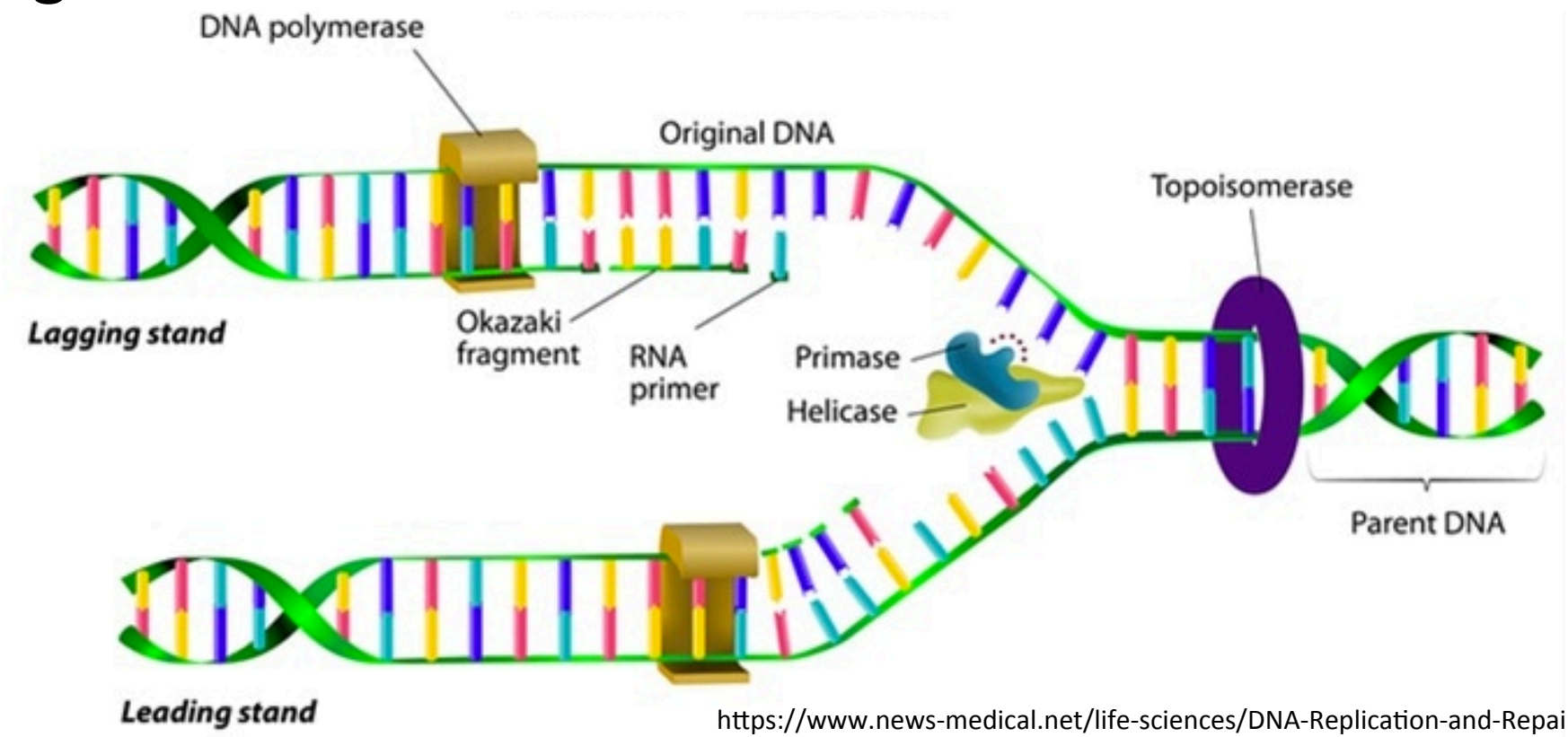
April 2.

[1] Pauling, L., and Corey, R. B., *Nature*, 171, 346 (1953); *Proc. U.S. Nat. Acad. Sci.*, 39, 84 (1953).
[2] Furberg, S., *Acta Chem. Scand.*, 6, 634 (1952).
[3] Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., *Biochim. et Biophys. Acta*, 9, 402 (1952).
[4] Wyatt, G. R., *J. Gen. Physiol.*, 36, 201 (1952).
[5] Astbury, W. T., Symp. Soc. Exp. Biol. 1, Nucleic Acid, 66 (Camb. Univ. Press, 1947).
[6] Wilkins, M. H. F., and Randall, J. T., *Biochim. et Biophys. Acta*, 10, 192 (1953).

Watson and Crick, 1953

DNA replication

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."



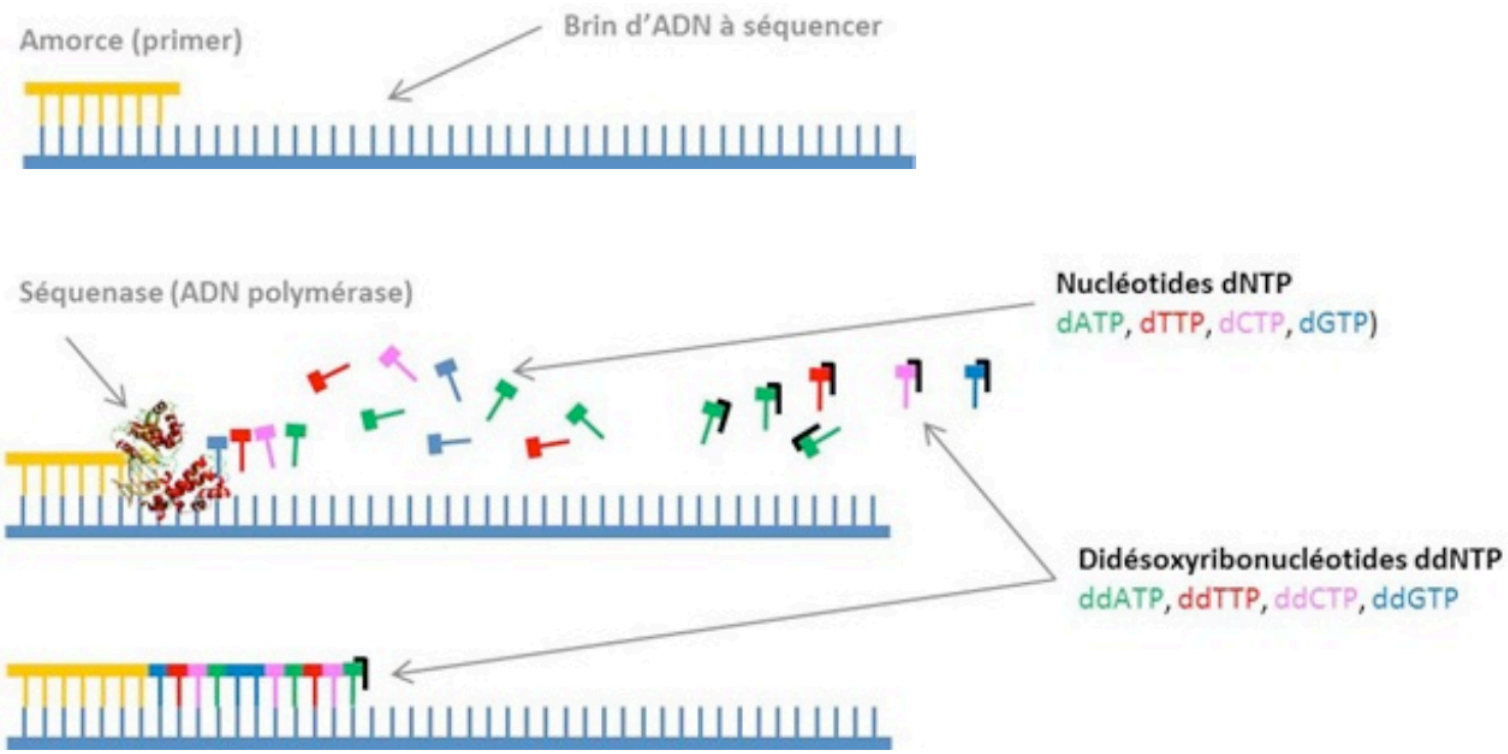
https://www.news-medical.net/life-sciences/DNA-Replication-and-Repair.aspx

1977: Sanger's Chain-Termination method

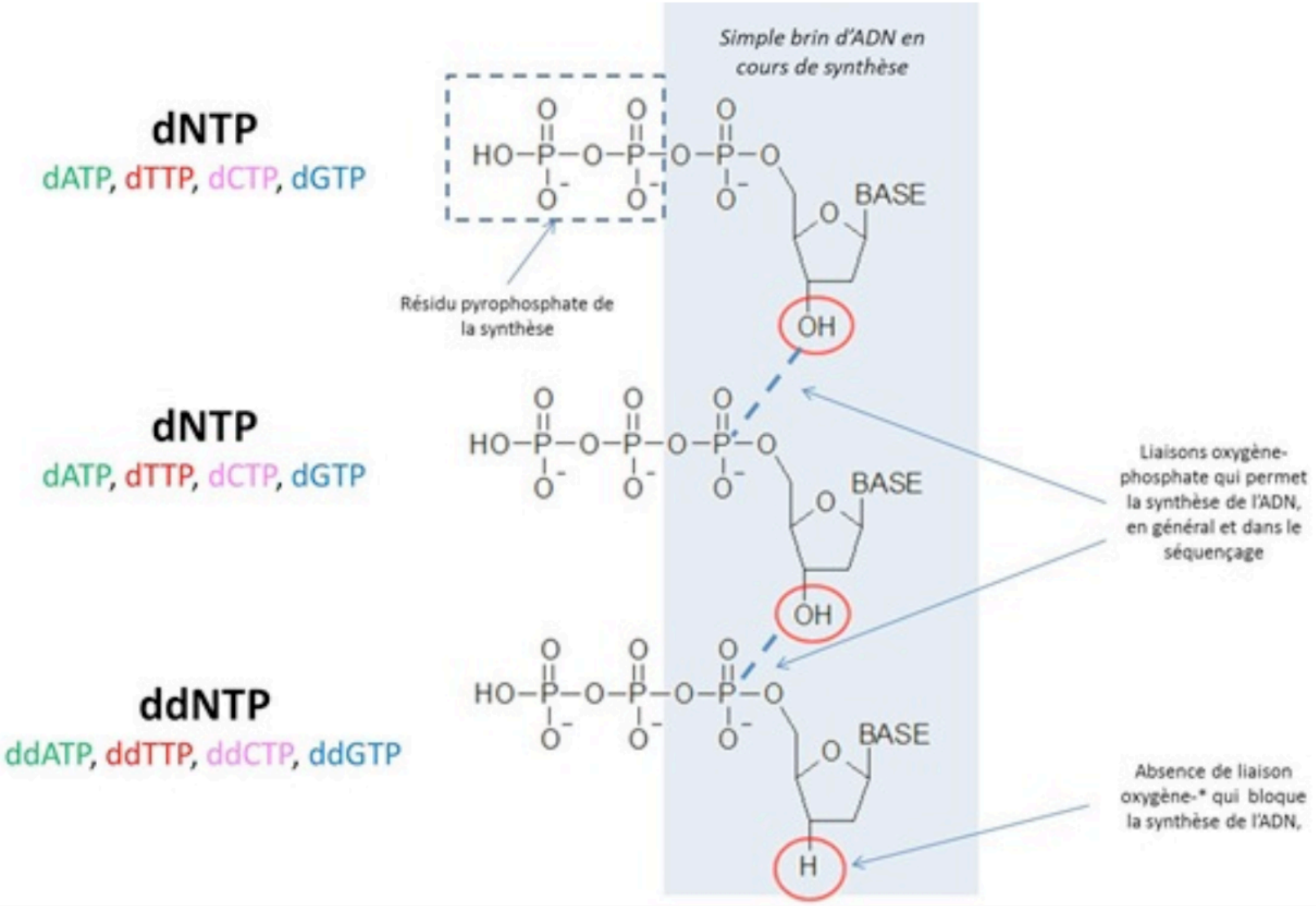
Denaturation & priming

Polymerization

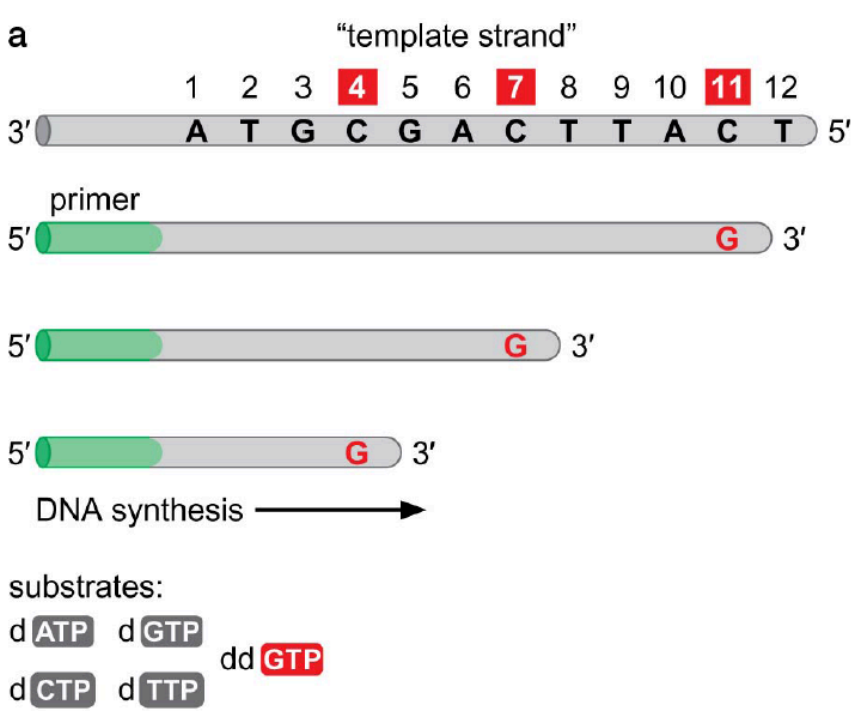
Arrest

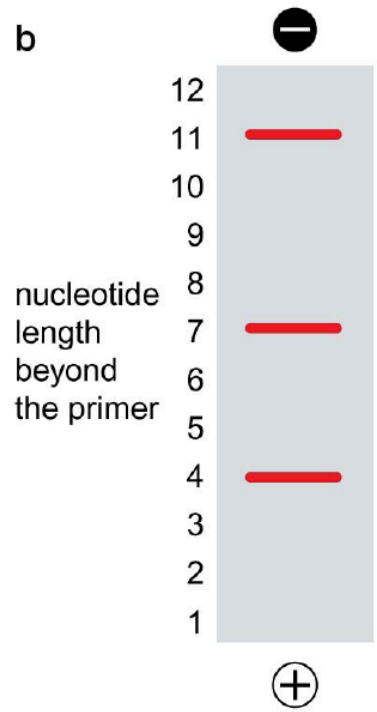


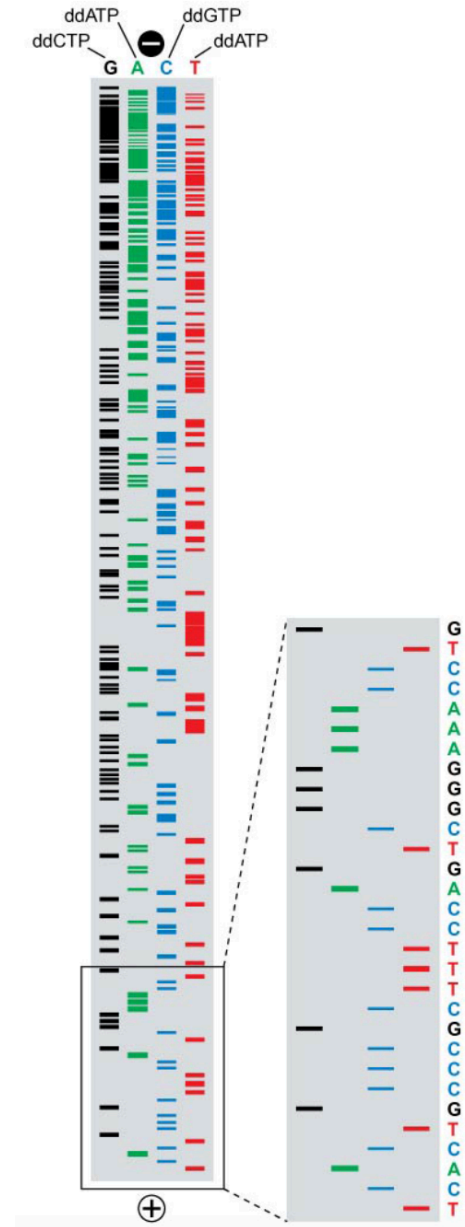
1977: Sanger's Chain-Termination method



1977: Sanger's Chain-Termination method

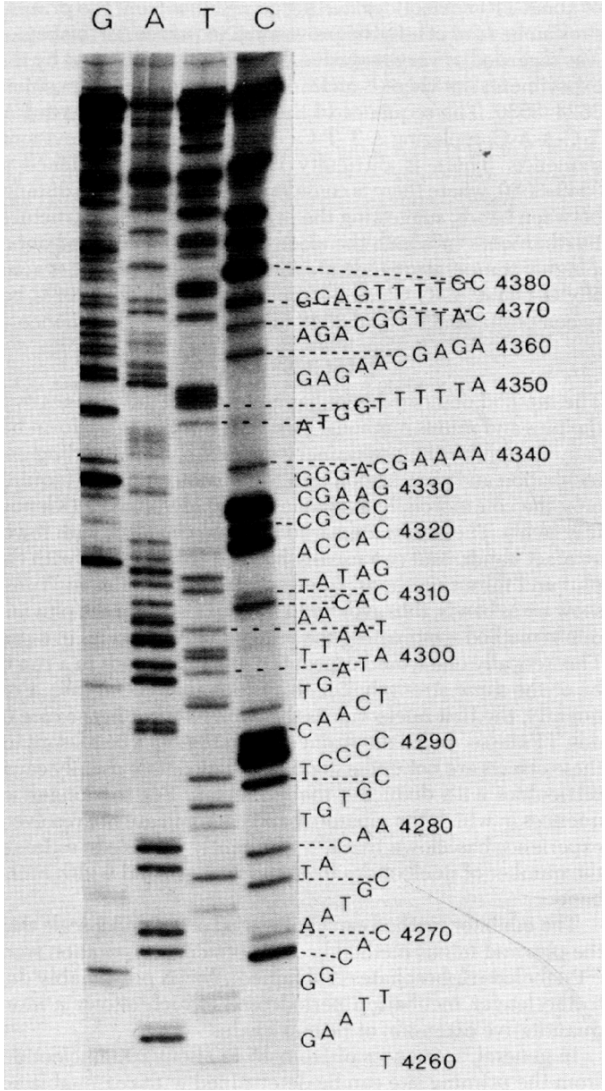






"sequencing by synthesis"

1977: Sanger's Chain-Termination method

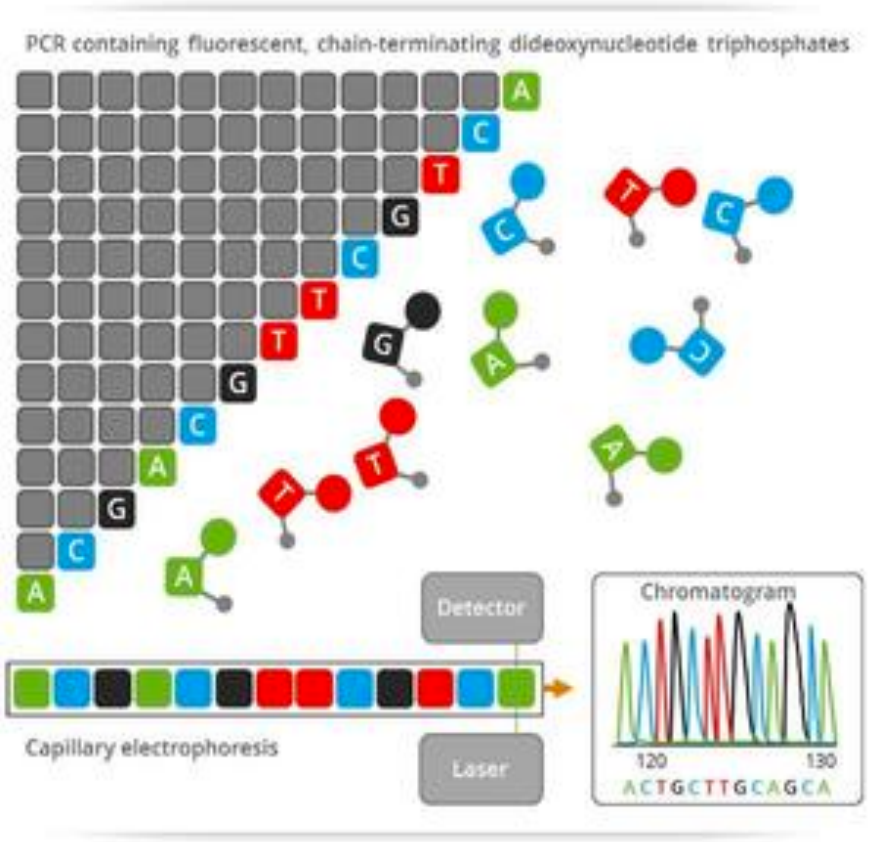


"The electrophoresis was on a 12% acrylamide gel at 40 mA for 14 hr."

Sanger *et al.* 1977

Improvements to the Sanger method

- Fluorophores instead of radio-labelling
- Capillary electrophoresis instead of acrylamide gels
- Discovery of more useful enzymes
- Automation



www.neogenomix.com/products/sanger-sequencing

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Format: Abstract

Send to

Cell. 1993 Mar 26;72(6):971-83.

A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group.

[No authors listed]

Abstract

The Huntington's disease (HD)
haplotype analysis of linkage d
A new gene, IT15, isolated usir
repeat that is expanded and un
observed on HD chromosomes
and 4p16.3 haplotypes. The (C
approximately 348 kd protein th
involves an unstable DNA segr
and myotonic dystrophy, acting

Comment in

Planting alfalfa and cloning the H

PMID: 8458085

[Indexed for MEDLINE]

Identification of the Cystic Fibrosis Gene: Cloning a
Riordan, John R;Rommens, Johanna M;Kerem, Bat-Sheva;Alon, I
Science; Sep 8, 1989; 245, 4922; Agricultural & Environmental S
pg. 1066

Identification of t
Cloning and
Comple

JOHN R. RIORDAN, JOHANNA M
RICHARD ROZMAHEL, ZBYSZE
NATASA PLAVSIC, JIA-LING CHOU,
FRANCIS S. COLLINS, LAP-CHEE TSUI

Overlapping complementary DNA clones were isolated from epithelial cell libraries with a genomic DNA segment containing a portion of the putative cystic fibrosis (CF) locus, which is on chromosome 7. Transcripts, approximately 6500 nucleotides in size, were detectable in the tissues affected in patients with CF. The predicted protein consists of two similar motifs, each with (i) a domain having properties consistent with membrane association and (ii) a domain believed to be involved in ATP (adenosine triphosphate) binding. A deletion of three

isolation of polypeptide comp
nel that mediates conductan
activated pathway and CF ha
biochemical defect in CF rem

Molecular cloning experim
large, contiguous segment o
scribed sequences from a reg
(7). These sequences were in
ability to detect conserved s
DNA hybridization and were
hybridization experiments, cl

MENU

nature
International journal of science

Article | Published: 24 January 1985

Complete nucleotide sequence of the AIDS virus, HTLV-III

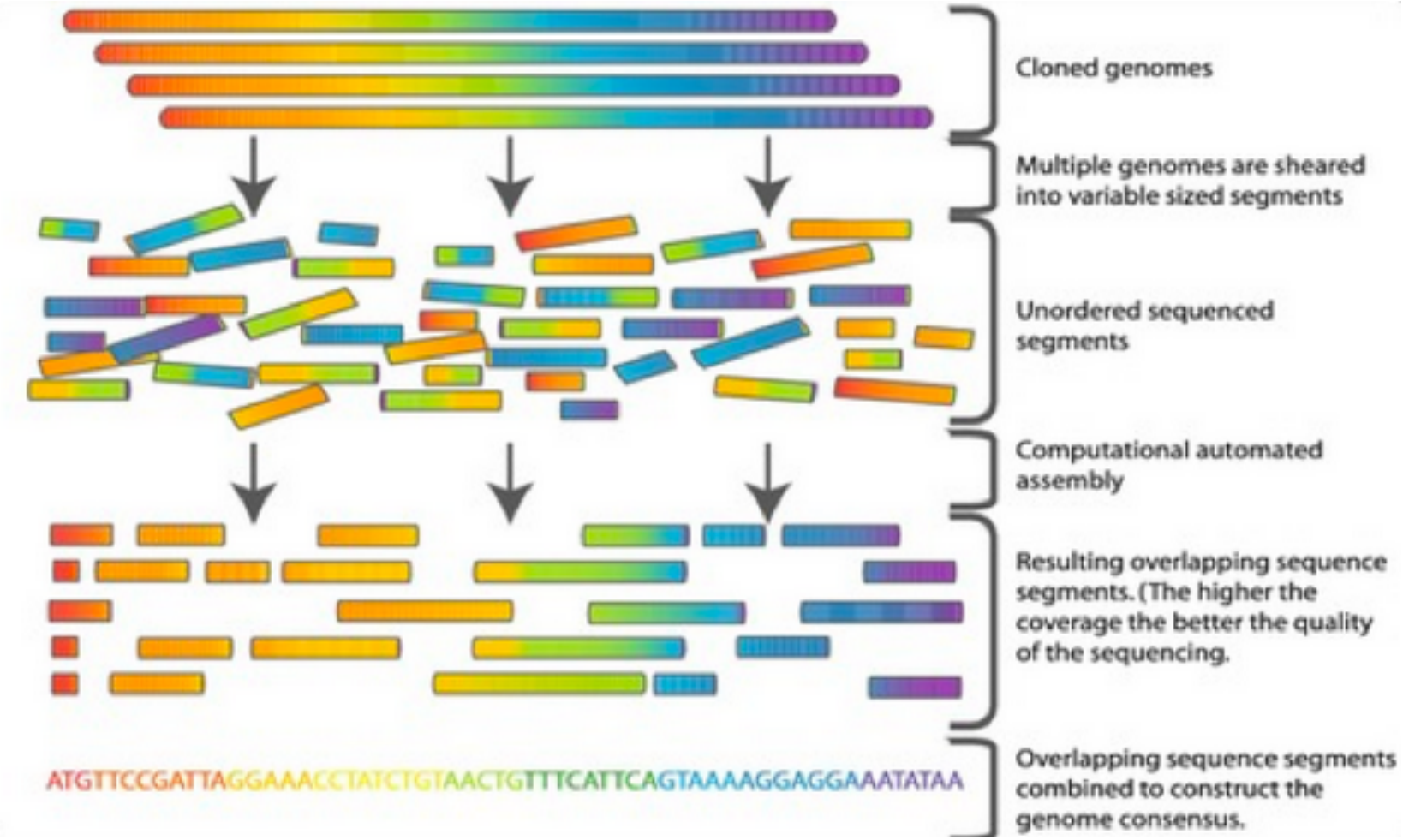
Lee Ratner, William Haseltine, Roberto Patarca, Kenneth J. Livak, Bruno Starcich, Steven F. Josep
Ellen R. Doran, J. Antoni Rafalski, Erik A. Whitehorn, Kirk Baumeister, Lucinda Ivanoff, Stephen R.
Petteway Jr, Mark L. Pearson, James A. Lautenberger, Takis S. Papas, John Ghrayeb, Nancy T. Cha
Robert C. Gallo & Flossie Wong-Staal

Nature **313**, 277–284 (24 January 1985) | Download Citation

Abstract

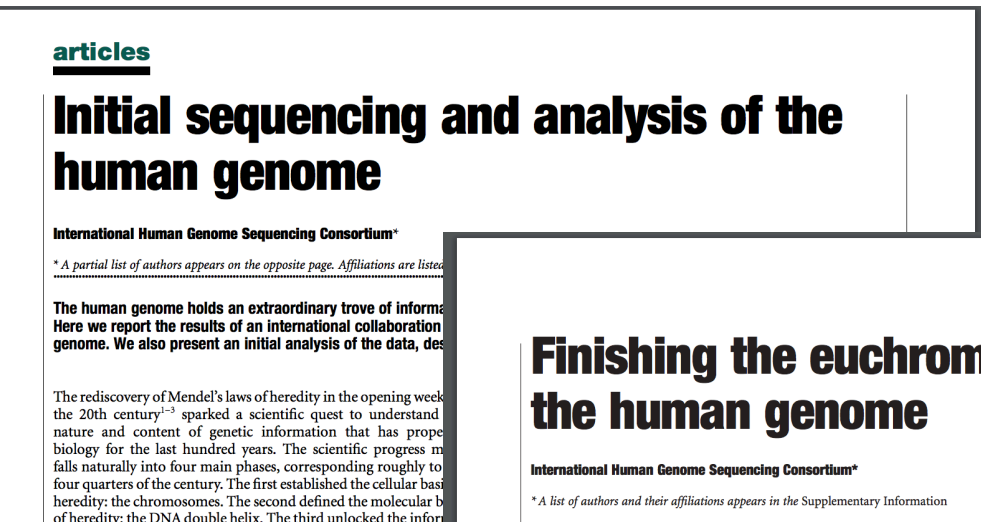
The complete nucleotide sequence of two human T-cell leukaemia typ
III (HTLV-III) proviral DNAs each have four long open reading frames,
the first two corresponding to the gag and pol genes. The fourth open
reading frame encodes two functional polypeptides, a large precursor
of the major envelope glycoprotein and a smaller protein derived from
the 3′-terminus long open reading frame analogous to the long open
reading frame (*lor*) product of HTLV-I and -II.

Shotgun Sequencing



Human Genome Project

- Initiated in 1990
  - Goal: finished genome by 2005
- Hundreds of labs in 18 countries
  - Sequences made available immediately after assembly
- Competitors: J. Craig Venter and Celera



articles

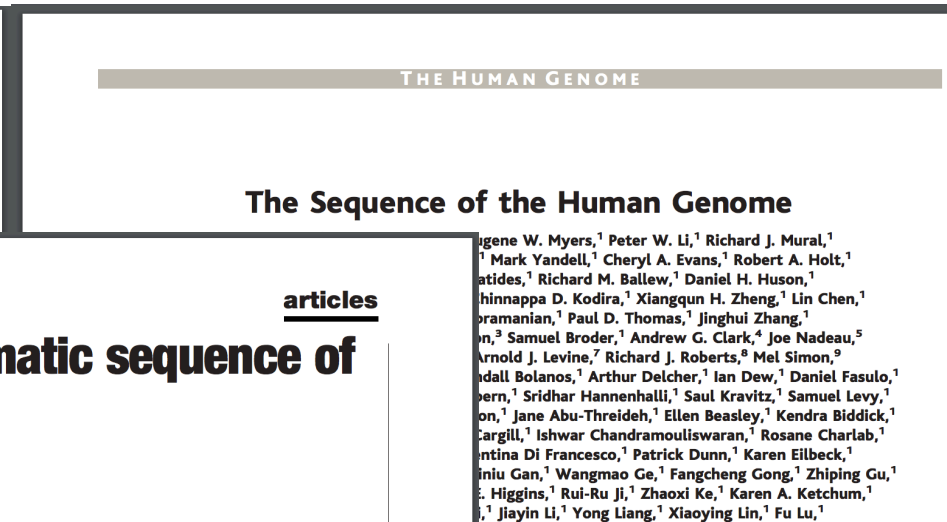
Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed

The human genome holds an extraordinary trove of informa
Here we report the results of an international collaboration
genome. We also present an initial analysis of the data, des

The rediscovery of Mendel's laws of heredity in the opening week
the 20th century[1–3] sparked a scientific quest to understand
nature and content of genetic information that has prope
biology for the last hundred years. The scientific progress m
falls naturally into four main phases, corresponding roughly to
four quarters of the century. The first established the cellular basi
heredity: the chromosomes. The second defined the molecular b
of heredity: the DNA double helix. The third unlocked the infor



THE HUMAN GENOME

The Sequence of the Human Genome

gene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
atides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
hinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]
bramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
n,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5]
rnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9]
ndall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1]
bern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1]
on,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1]
Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1]
ntina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1]
iniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1]
. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1]
,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]

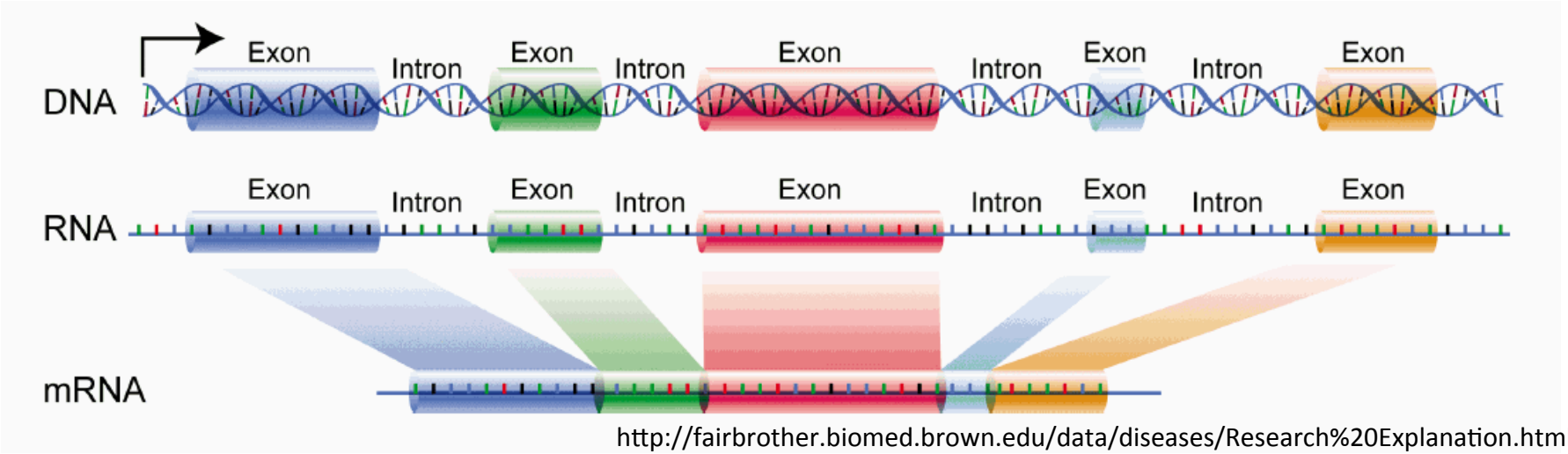
articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

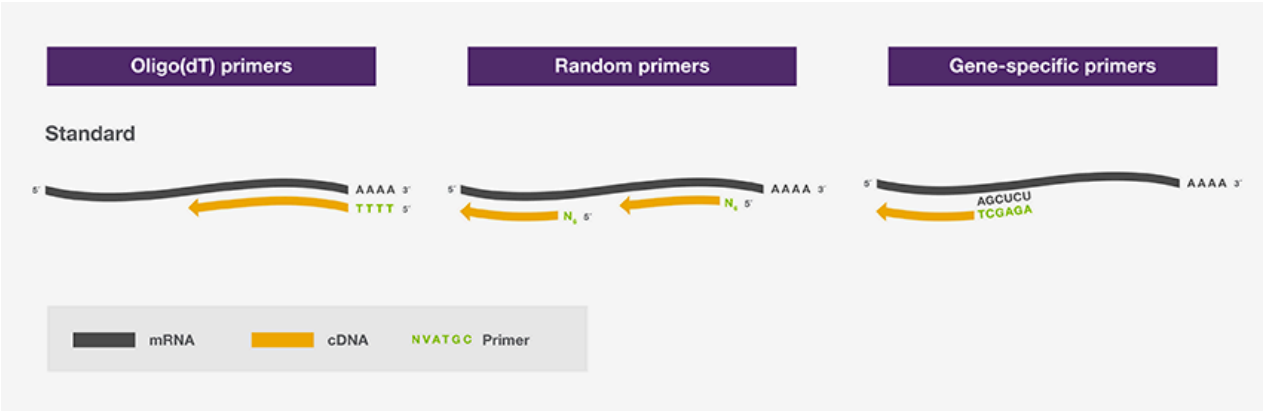
* A list of authors and their affiliations appears in the Supplementary Information

Sequencing RNA



http://fairbrother.biomed.brown.edu/data/diseases/Research%20Explanation.html

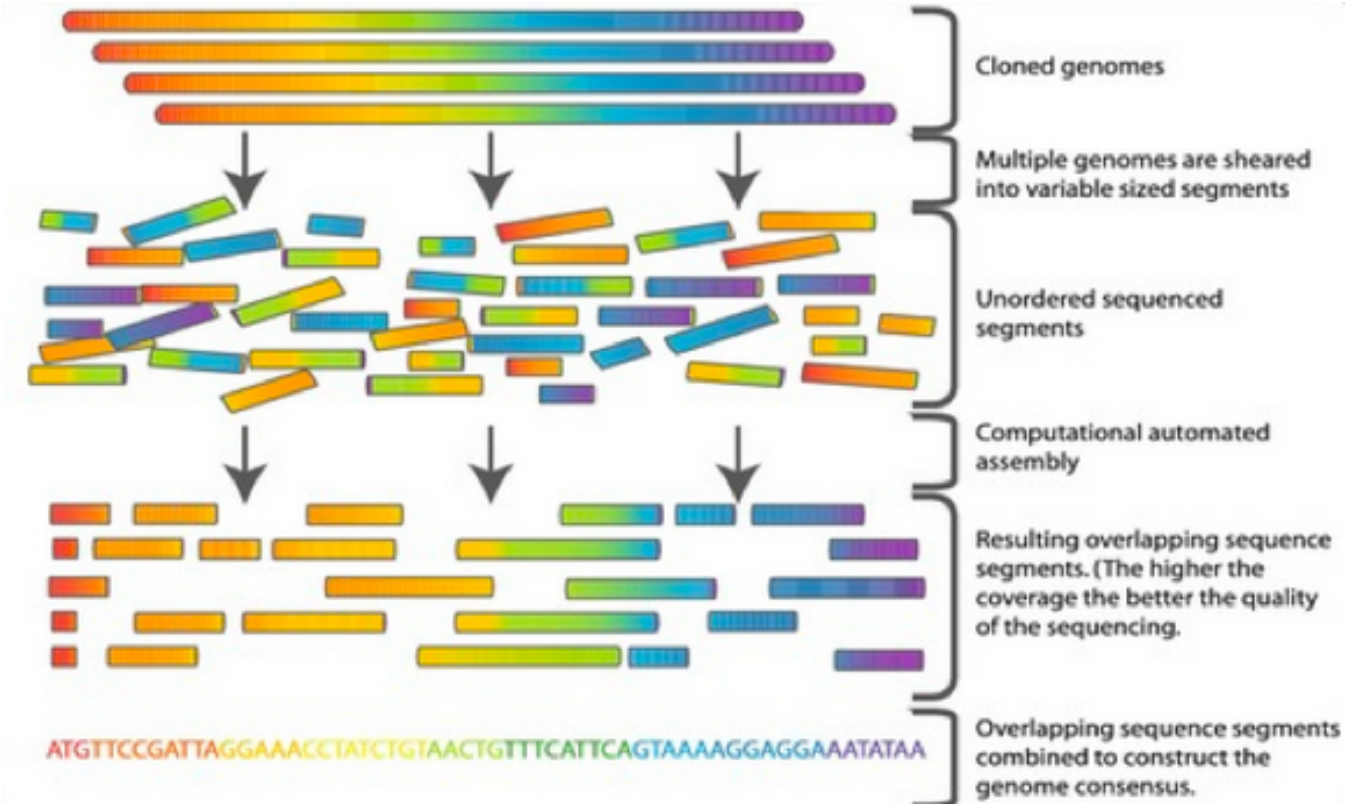
Reverse transcription to cDNA is (almost) always necessary.



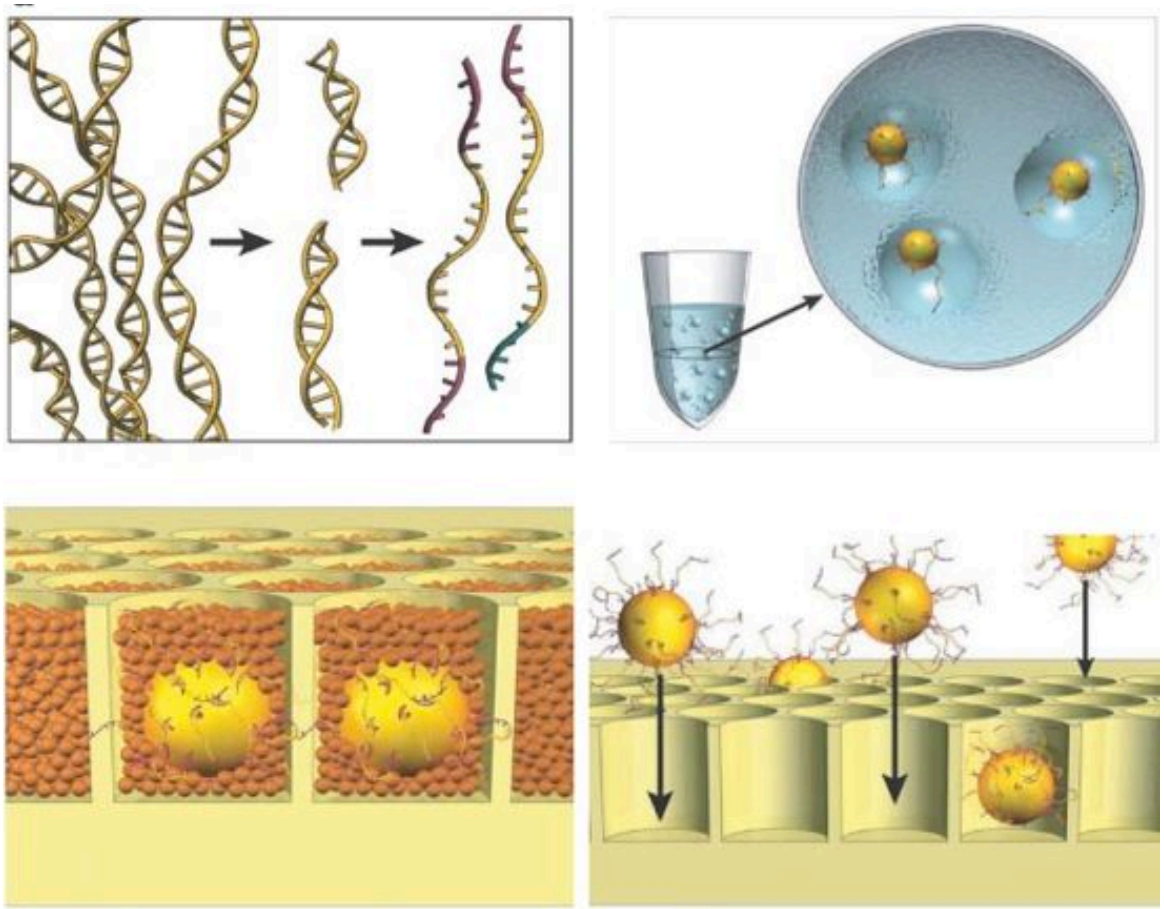
https://www.thermofisher.com/be/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/rt-education/reverse-transcription-setup.html

Second Generation Sequencing

- AKA "next generation sequencing" or NGS
- Massively parallel sequencing
- Short read length

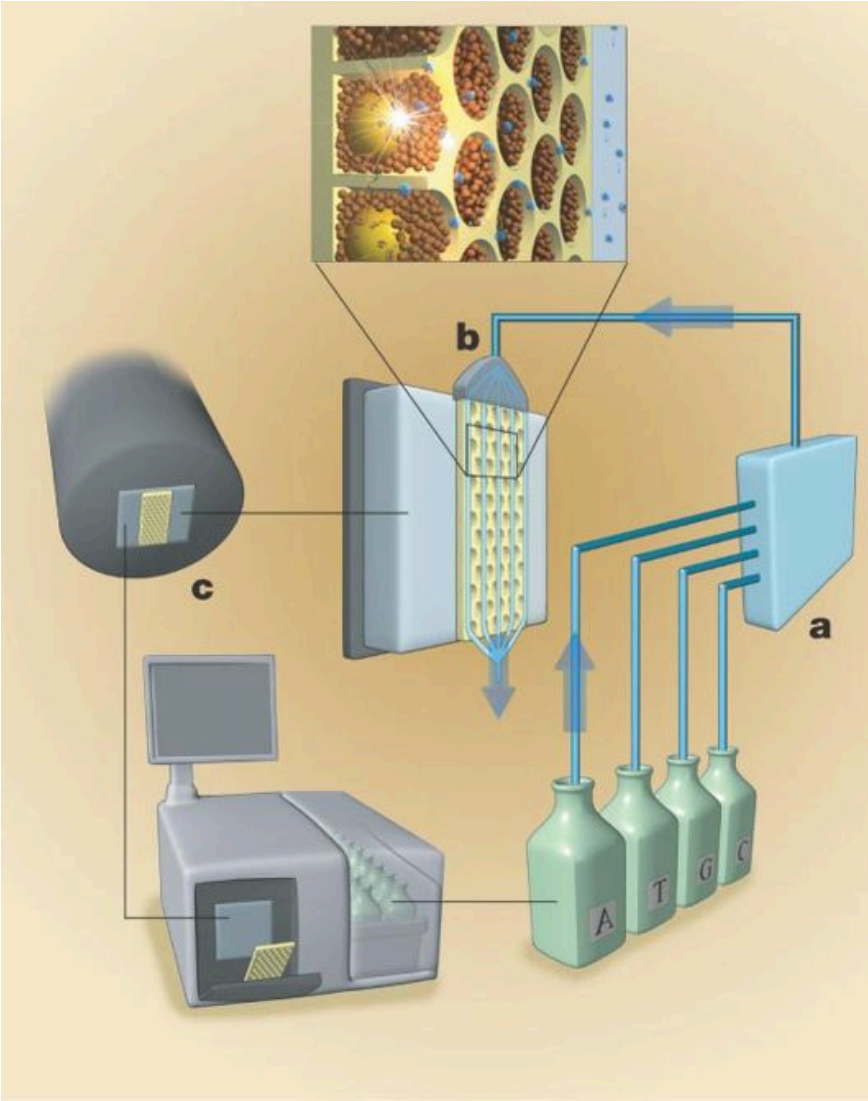


~2005: 454 Pyrosequencing



Margulies *et al.* 2005

454 Pyrosequencing



Margulies *et al.* 2005

Genome Sequencing with 454

"Here we report the DNA sequence of a diploid genome of a single individual, James D. Watson, sequenced to 7.4-fold redundancy in **two months** using massively parallel sequencing in picolitre-size reaction vessels."

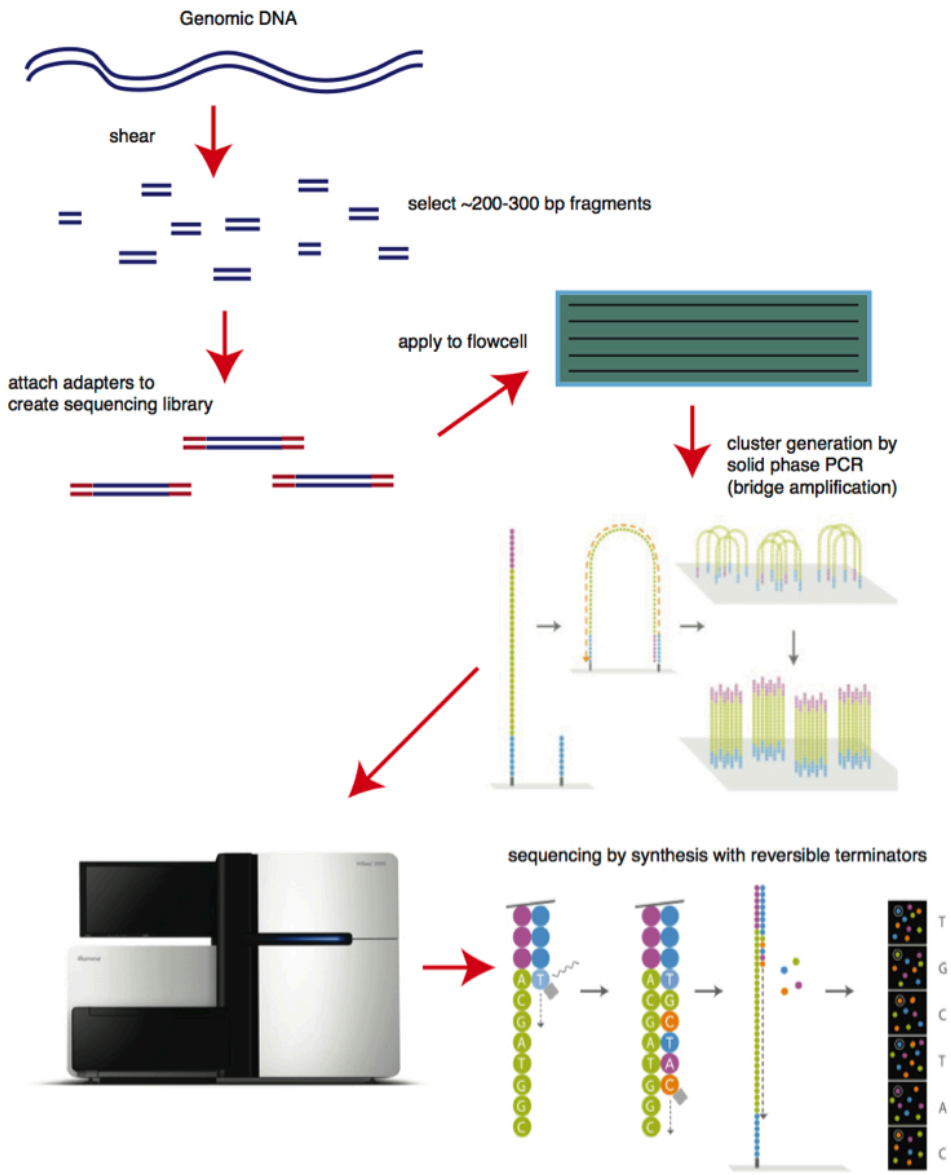
nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2],

~2006: Solexa/Illumina



https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

2012

2016

Variant analysis

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rar such as motif-disrupting changes in transcription-factor-binding sites. This accessible single nucleotide polymorphisms at a frequency of 1% in related pop low-frequency variants in individuals from diverse, including admixed, popu

Recent efforts to map human genetic variation by sequencing exomes[1] and whole genomes[2–4] have characterized the vast majority of common single nucleotide polymorphisms (SNPs) and many structural variants across the genome. However, although more than 95% of common (>5% frequency) variants were discovered in the pilot phase of the 1000 Genomes Project, lower-frequency variants, particularly those outside the coding exome, remain poorly characterized. Low-frequency variants are enriched for potentially functional mutations, for example, protein-changing variants, under weak purifying selection[1,5,6]. Furthermore, because low-frequency variants tend to be recent in origin, they exhibit increased levels of population differentiation[6–8]. Characterizing such variants, for both point mutations and structural changes, across a range of populations is thus likely to identify many variants of functional importance and is crucial for interpreting individual genome those private to fan

We now report on the genomes of 1,092 individuals sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and the Americas (Supplementary Figs 1 and 2), analysed through a combination of low-coverage (2–6×) whole-genome sequence data, targeted deep (50–100×) exome sequence data and dense SNP genotype data (Table 1 and Supplementary Tables 1–3). This design was shown by the pilot phase[2] to be powerful and cost-effective in discovering and genotyping all but the rarest SNP and short insertion and deletion (indel) variants. Here, the approach was augmented with statistical methods for selecting higher quality variant calls from candidates obtained using multiple algorithms, and to integrate SNP, indel and larger structural variants within a single framework (see

Table 1 | Summary of 1000 Genomes Project phase I data

	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K
Novelty rate†	58%	77%	50%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K	199/293/6.3 K
Average no. SNPs per sample	3.60 M	105 K	24.0 K
Indels			
No. sites overall	1.38 M	59 K	1,867
Novelty rate†	62%	73%	54%
No. inframe/frameshift	NA	19/14	719/1,066
Average no. indels per sample	344 K	13 K	440
Genotyped large deletions			
No. sites overall	13.8 K	432	847
Novelty rate†	54%	54%	50%
Average no. variants per sample	717	26	39

NA, not applicable.
* Autosomal genes only.
† Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).

*Lists of participants and their affiliations appear at the end of the paper.

thebmj

BMJ 2016;353:i1757 doi: 10.1136/bmj.i1757 (Published 13 April 2016)

FEATURE

GENETIC RESEARCH

The 100 000 Genomes Project

Part research project, part commercial stimulus, this enormous sequencing programme could usher genomic medicine into mainstream use, **Mark Peplow** reports

Mark Peplow *freelance journalist, Cambridge*

is now ramping up into high gear. Overseen by Genomics England, it is one of the biggest whole genome sequencing projects in the world. And it is working to a breathtaking timetable: most of these genomes will be sequenced by the end of next year.

The genetic material will come from patients with rare diseases or common cancers and their families (box 1). By identifying any genetic anomalies, and linking them to participants' medical histories for the rest of their lives, the project aims to build up a unique database for treatment and research. "It will allow us to find things in the data that we might not notice in ordinary clinical care," says Caulfield. That should offer better diagnoses and more targeted therapies. It also gives scientists a treasure trove of information that could help to develop more effective drugs.

That remit is impressive enough. But the project's broader goals are to kickstart a national genomics industry and make the UK the first country to routinely use DNA sequencing in mainstream healthcare. "If we get this right, our ambition is to see new treatments, new diagnostics, coming to patients in the UK first," says Caulfield.

Clinical potential

The project is already having clinical impact among people with rare diseases, with the first child participants receiving a genetic diagnosis in January. There are about 7000 known rare diseases, and roughly 1 in 17 people (about three million in the UK) are affected at some point in their lives.[1] "Collectively, the burden is high," says Caulfield. "They are a huge cause of disability, and the toll on individuals is huge."

More than 80% of rare diseases are suspected to have a genetic component. But their rarity makes diagnosis a huge challenge,

Beverly Searle, chief executive of Unique, the rare chromosome disorder support group (www.rarechromo.co.uk). "That's why it's so important to have these projects where you gather large datasets." With tens of thousands of genomes from patients with rare diseases, it becomes much more likely to find a statistically robust association between genetic variants and a particular disease.

Once that link is established, the experiences of those who share the same genetic anomalies can be compared to predict how a particular patient's condition might develop in the future and which treatments are likely to be more effective. Not only could this improve clinical outcomes, it could also save time and money. "It's important that expectations are managed—not every family will get a diagnosis," cautions Searle. "But we've got the potential for one test to give you an answer."

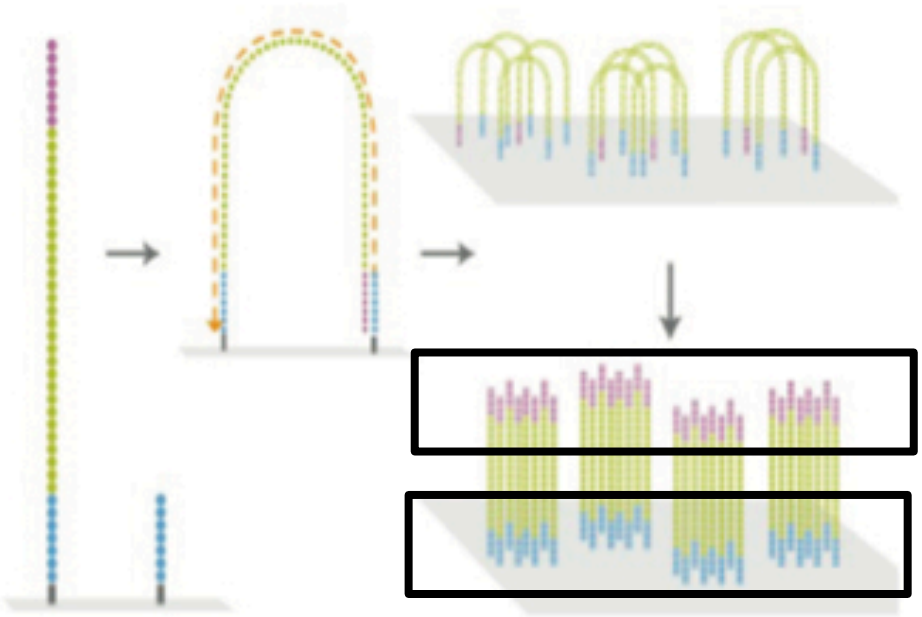
The other arm of the project focuses on common cancers, including those in the lung, breast, colon, prostate, and ovary, where a genetic diagnosis could affect treatment options. About half of melanomas are caused by a mutation in the *BRAF* gene, for example, and these can be treated with a drug that specifically targets the BRAF protein. "A mutation can help to predict a medicine's effectiveness," says Caulfield.

Collecting and using the data

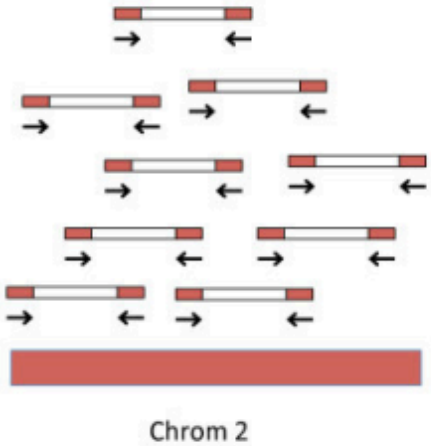
Most of the project's participants arrive via one of the 13 NHS Genomic Medicine Centres that were established last year around England. People give a small blood sample, and (if they have cancer) a small piece of their tumour, which can have a substantially different genome.

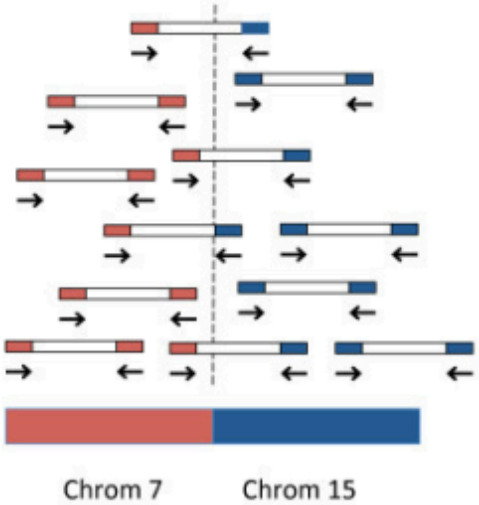
The project has already sequenced more than 7000 genomes and is recruiting more than 200 patients with rare diseases per week. But the pace of sequencing will quicken in the coming months, when a dedicated facility at the Wellcome Genome Campus in Hinxton, Cambridgeshire, opens. The American company Illumina is setting up a world class sequencing facility there, stuffed with machines that can read an entire genome in

Paired-end sequencing



https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/





Vollan and Caldas 2011

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zhe Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis L Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bo S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health we describe the Illumina-based metagenomic sequencing, microbial genes, derived from 576.7 gigabases of sequence ~150 times larger than the human gene complement, cont microbial genes of the cohort and probably includes a large genes are largely shared among individuals of the cohort. cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in term functions present in all individuals and most bacteria, respectively.

Metagenomics

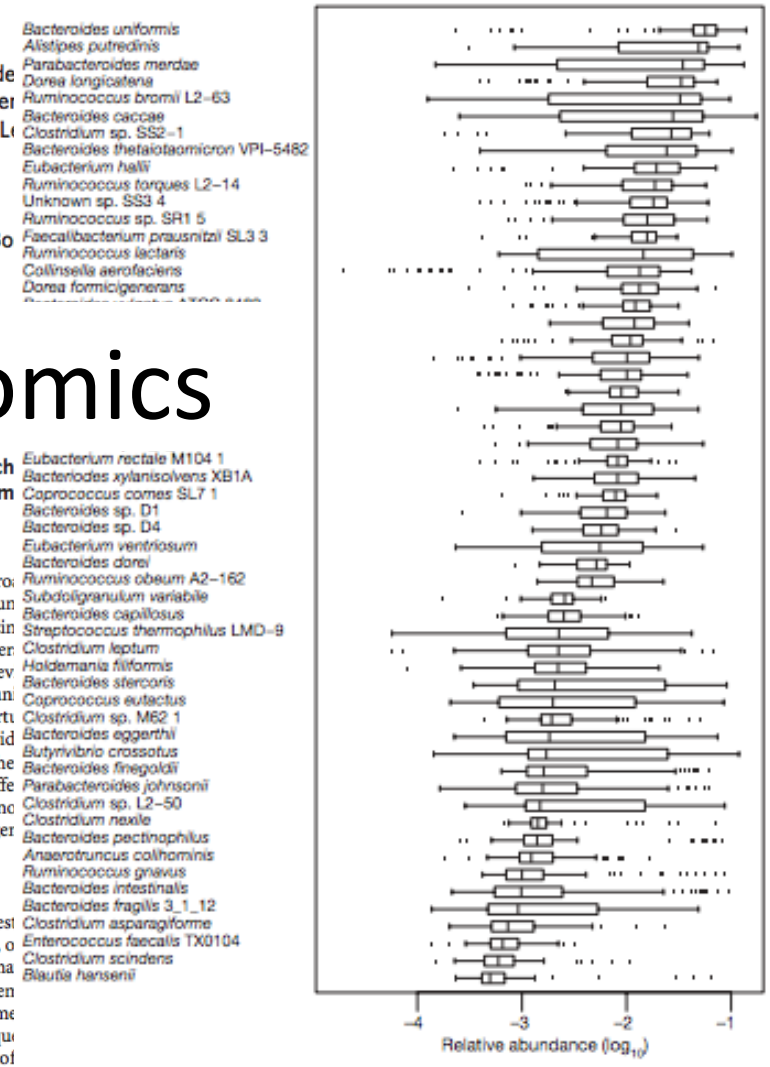
It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome[1]. The majority of microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial for human life[2,3]. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity[4–8].

To understand and exploit the impact of the gut microbes on human health and well-being it is necessary to decipher the content, diversity and functioning of the microbial gut community. 16S ribosomal RNA gene (rRNA) sequence-based methods[9] revealed that two bacterial divisions, the Bacteroidetes and the Firmicutes, constitute over 90% of the known phylogenetic categories and dominate the distal gut microbiota[10]. Studies also showed substantial diversity of the gut microbiome between healthy individuals[4,8,10,11]. Although this difference is especially marked among infants[12], later in life the gut microbiome converges to more similar phyla.

Metagenomic sequencing represents a powerful alternative to rRNA sequencing for analysing complex microbial communities[13–15]. Applied to the human gut, such studies have already generated some 3 gigabases (Gb) of microbial sequence from faecal samples of 33 individuals from the United States or Japan[8,16,17]. To get a bro overview of the human gut microbial genes we used the Illu Genome Analyser (GA) technology to carry out deep sequencin total DNA from faecal samples of 124 European adults. We gener 576.7 Gb of sequence, almost 200 times more than in all prev studies, assembled it into contigs and predicted 3.3 million un open reading frames (ORFs). This gene catalogue contains virtu all of the prevalent gut microbial genes in our cohort, provid broad view of the functions important for bacterial life in the and indicates that many bacterial species are shared by diffe individuals. Our results also show that short-read metagen sequencing can be used for global characterization of the ge potential of ecologically complex environments.

Metagenomic sequencing of gut microbiomes

As part of the MetaHIT (Metagenomics of the Human Intes Tract) project, we collected faecal specimens from 124 healthy, o weight and obese individual human adults, as well as inflamma bowel disease (IBD) patients, from Denmark and Spain (Supplen tary Table 1). Total DNA was extracted from the faecal specime and an average of 4.5 Gb (ranging between 2 and 7.3 Gb) of sequ was generated for each sample, allowing us to capture most of



Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,[1]* Ali Mortazavi,[2]* Richard M. Myers,[1]† Barbara Wold[2,3]†

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position [±50 base pairs (bp)], which facilitated our finding motifs and allowed us to ident
data also have high sensitivity and spec
area ≥ 0.96] and statistical confidence
candidate interactions. These include k
pancreatic islet cell development.

ChIP-Seq

RIP-Seq

…

Although much is known about t
tion factor binding and action a
genes, far less is known about
position and function of entire fact
interactomes, especially for organisms w
genomes. Now that human, mouse, a
large genomes have been sequence
possible, in principle, to measure h
transcription factor is deployed across t
genome for a given cell type and phys
condition. Such measurements are important for systems-level studies because they provide a global map of candidate gene network input connections. These direct physical interactions between transcription factors or cofactors and the

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305–5120, USA. [2]Biology Division, California Institute of Technology, Pasadena, CA 91125, USA. [3]California Institute of Technology Beckman Institute, Pasadena, CA 91125, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: woldb@its.caltech.edu (B.W.); myers@shgc.stanford.edu (R.M.M.)

pleteness, and high binding-site resolution. These data-quality and depth issues dictate whether primary gene network structure can be inferred with reasonable certainty and comprehensiveness, and how effectively the data can be used to discover binding-site motifs by computational methods. For these purposes, statistical robustness, sampling depth across the genome, absolute signal and signal-to-noise ratio must be good enough to detect nearly all in vivo binding locations for a regulator with minimal inclusion of false-positives. A further challenge in genomes large or small is to map factor-binding sites with high positional resolution. In addition to making computational discovery of binding motifs feasible, this dictates the quality of regulatory site annotation relative to other gene anatomy landmarks, such as transcription start sites, enhancers, introns and exons, and conserved noncoding features (*2*). Finally, if high-quality protein-DNA interactome measurements can be performed routinely and at reasonable cost, it will open the way to detailed studies of interactome dynamics in response to specific signaling stimuli or genetic mutations. To address these issues, we turned to ultrahigh-throughput DNA sequencing to gain sampling power and applied size selection on immuno-enriched DNA to enhance positional resolution.

shown here differs
hIP methods such as
Pchip (*1*); ChIPSAGE
t (*4*) in design, data
design is simple (Fig.
ChIPPet, it involves no
on. Unlike microarray
of single-copy sites in
for ChIPSeq assay (*5*),
ed to be array features.
with similar complete-
le microarray design, a
liding window design
es per array would be
portion of the human
Seq counts sequences
and so avoids constraints imposed by array hybridization chemistry, such as base composition constraints related to T_m, the temperature at which 50% of double-stranded DNA or DNA-RNA hybrids is denatured; cross-hybridization; and secondary structure interference. Finally, ChIPSeq is feasible for any sequenced genome, rather than being restricted to species for which whole-genome tiling arrays have been produced.

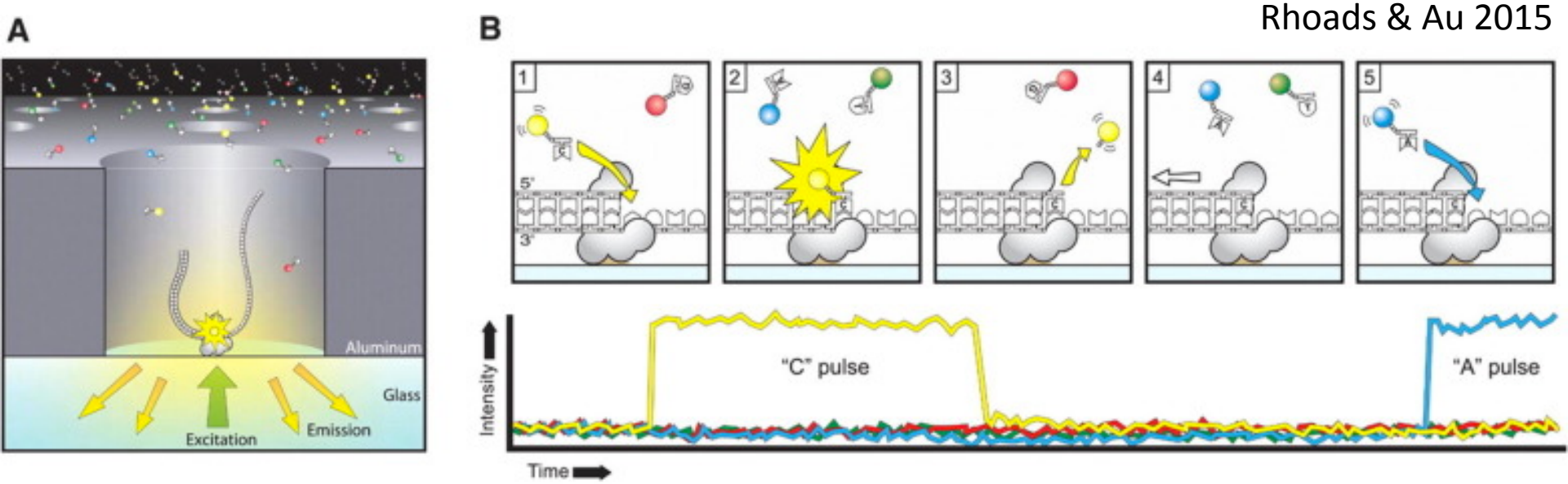
ChIPSeq illustrates the power of new sequencing platforms, such as those from Solexa/Illumina and 454, to perform sequence census counting assays. The generic task in these applications is to identify and quantify the molecular

Single molecule sequencing

- Sometimes called “third generation sequencing”
- Long read lengths (like Sanger sequencing) and high throughput (like second-generation sequencing)

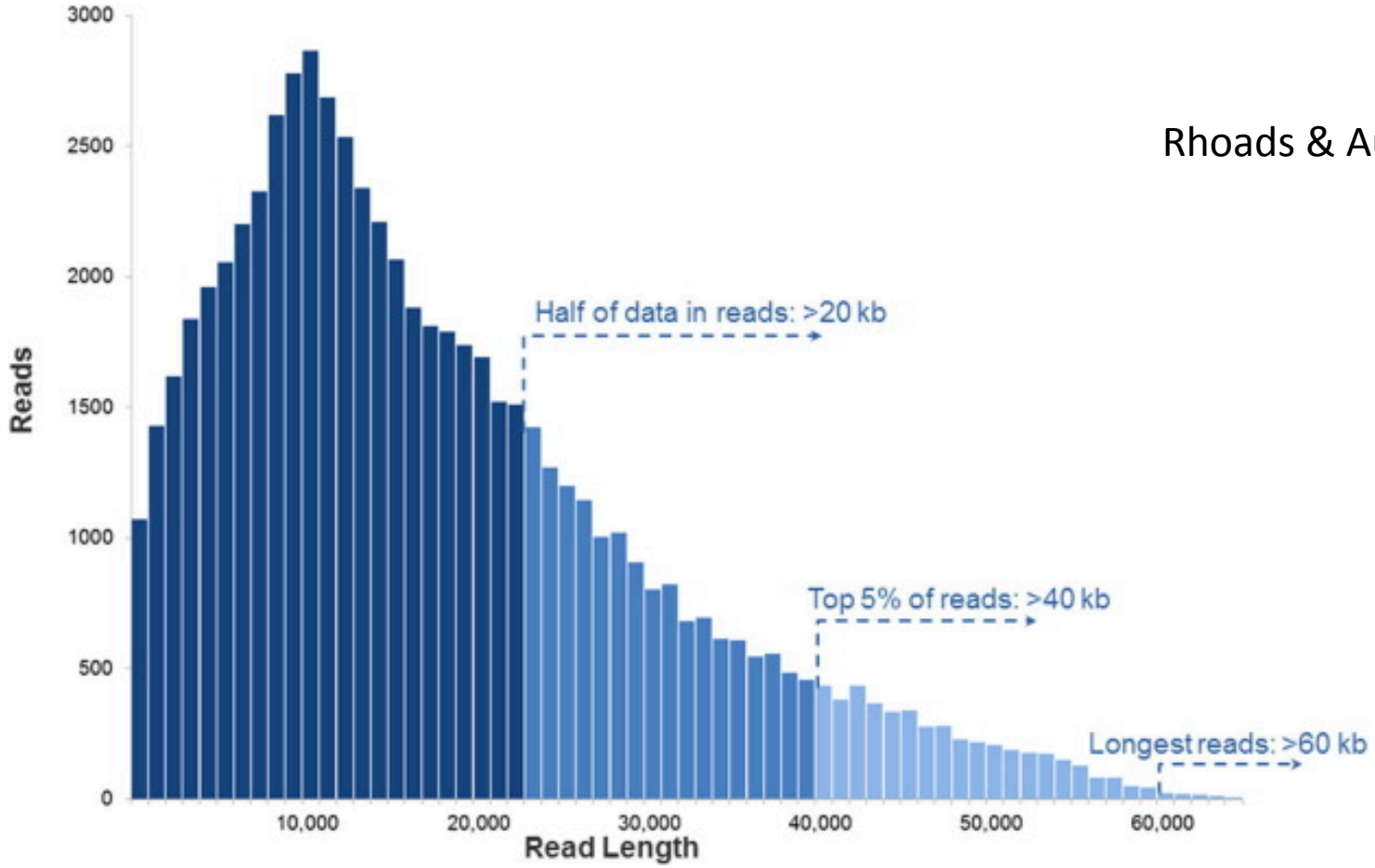
~2013: PacBio SMRT sequencing

Rhoads & Au 2015



Similar to 454 sequencing, but optics can detect the flash of a **single** dNTP being incorporated

PacBio SMRT sequencing



Rhoads & Au 2015

Reads are **much** longer than Illumina reads

Wellcome Open Research

Wellcome Open Research 2017, 2:42 Last updated: 09 MAY 2018

Check for updates

RESEARCH ARTICLE

An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion [version 1; referees: 2 approved]

Erica M Pasini[1], Ulrike Böhme[2], Gavin G. Rutledge[2], Annemarie Voorberg-Van der Wel[1], Mandy Sanders[2], Matt Berriman[2], Clemens HM Kocken[1], Thomas D. Otto[2]

[1]Biomedical Primate Research Centre, Rijswijk, Lange Kleiweg 161, 2288GJ Rijswijk, Netherlands
[2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

v1 **First published:** 16 Jun 2017, **2**:42 (doi: 10.12688/wellcomeopenres.11864.1)
Latest published: 16 Jun 2017, **2**:42 (doi: 10.12688/wellcomeopenres.11864.1)

Abstract

Background: *Plasmodium cynomolgi*, a non-human primate malaria parasite species, has been an important model parasite since its discovery in 1907. Similarities in the biology of *P. cynomolgi* to the closely related, but less tractable, human malaria parasite *P. vivax* make it the model parasite of choice for liver biology and vaccine studies pertinent to *P. vivax* malaria. Molecular and genome-scale studies of *P. cynomolgi* have relied on the current reference genome sequence, which remains highly fragmented with 1,649 unassigned scaffolds and little representation of the subtelomeres.
Methods: Using long-read sequence data (Pacific Biosciences SMRT technology), we assembled and annotated a new reference genome sequence, PcyM, sourced from an Indian rhesus monkey. We compare the newly assembled genome sequence with those of several other *Plasmodium* species, including a re-annotated *P. coatneyi* assembly.
Results: The new PcyM genome assembly is of significantly higher quality than the existing reference, comprising only 56 pieces, no gaps and an improved average gene length. Detailed manual curation has ensured a comprehensive annotation of the genome with 6,632 genes, nearly 1,000 more than previously attributed to *P. cynomolgi*. The new assembly also has an improved representation of the subtelomeric regions, which account for nearly 40% of the sequence. Within the subtelomeres, we identified more than 1300 *Plasmodium* interspersed repeat (*pir*) genes, as well as a striking expansion of 36 methyltransferase pseudogenes that originated from a single copy on chromosome 9.
Conclusions: The manually curated PcyM reference genome sequence is an important new resource for the malaria research community. The high quality and contiguity of the data have enabled the discovery of a novel expansion of methyltransferase in the subtelomeres, and illustrates the new comparative genomics capabilities that are being unlocked by complete reference genomes.

Keywords

P. cynomolgi, PacBio assembly, P. coatneyi, methyltransferase

Open Peer Review

Referee Status: ✓ ✓

	Invited Refer
	1
version 1 published 16 Jun 2017	✓ report

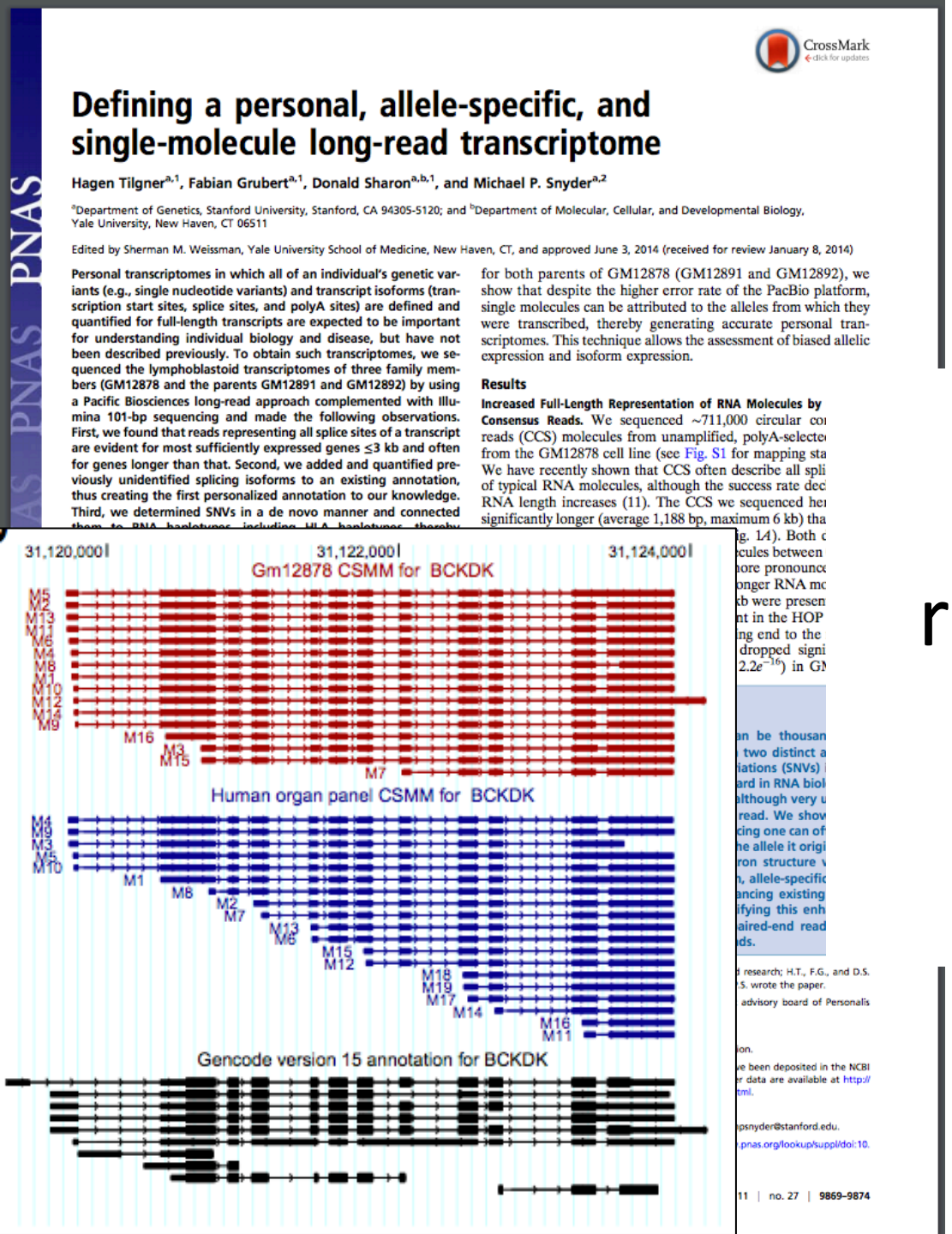
1 **Aaron R. Jex**, Walter and Eliza Institute of Medical Research, A

2 **Richárd Bártfai**, Radboud Univ Netherlands

Discuss this article

Comments (0)

Better resolution of genomic sequence (especially in repeat/ expansion regions)



Defining a personal, allele-specific, and single-molecule long-read transcriptome

Hagen Tilgner[a,1], Fabian Grubert[a,1], Donald Sharon[a,b,1], and Michael P. Snyder[a,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305-5120; and [b]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06511

Edited by Sherman M. Weissman, Yale University School of Medicine, New Haven, CT, and approved June 3, 2014 (received for review January 8, 2014)

Personal transcriptomes in which all of an individual's genetic variants (e.g., single nucleotide variants) and transcript isoforms (transcription start sites, splice sites, and polyA sites) are defined and quantified for full-length transcripts are expected to be important for understanding individual biology and disease, but have not been described previously. To obtain such transcriptomes, we sequenced the lymphoblastoid transcriptomes of three family members (GM12878 and the parents GM12891 and GM12892) by using a Pacific Biosciences long-read approach complemented with Illumina 101-bp sequencing and made the following observations. First, we found that reads representing all splice sites of a transcript are evident for most sufficiently expressed genes ≤3 kb and often for genes longer than that. Second, we added and quantified previously unidentified splicing isoforms to an existing annotation, thus creating the first personalized annotation to our knowledge. Third, we determined SNVs in a de novo manner and connected

for both parents of GM12878 (GM12891 and GM12892), we show that despite the higher error rate of the PacBio platform, single molecules can be attributed to the alleles from which they were transcribed, thereby generating accurate personal transcriptomes. This technique allows the assessment of biased allelic expression and isoform expression.

Results

Increased Full-Length Representation of RNA Molecules by Consensus Reads. We sequenced ~711,000 circular co reads (CCS) molecules from unamplified, polyA-selecte from the GM12878 cell line (see Fig. S1 for mapping sta We have recently shown that CCS often describe all spli of typical RNA molecules, although the success rate dec RNA length increases (11). The CCS we sequenced he significantly longer (average 1,188 bp, maximum 6 kb) tha g. 1A). Both c cules between hore pronounc onger RNA mo kb were presen nt in the HOP ng end to the dropped signi $2.2e^{-16}$) in GM

an be thousan two distinct a iations (SNVs) ard in RNA biol although very u read. We show cing one can of he allele it origi ron structure allele-specific ancing existing ifying this enh aired-end read ds.

d research; H.T., F.G., and D.S. S. wrote the paper. advisory board of Personalis

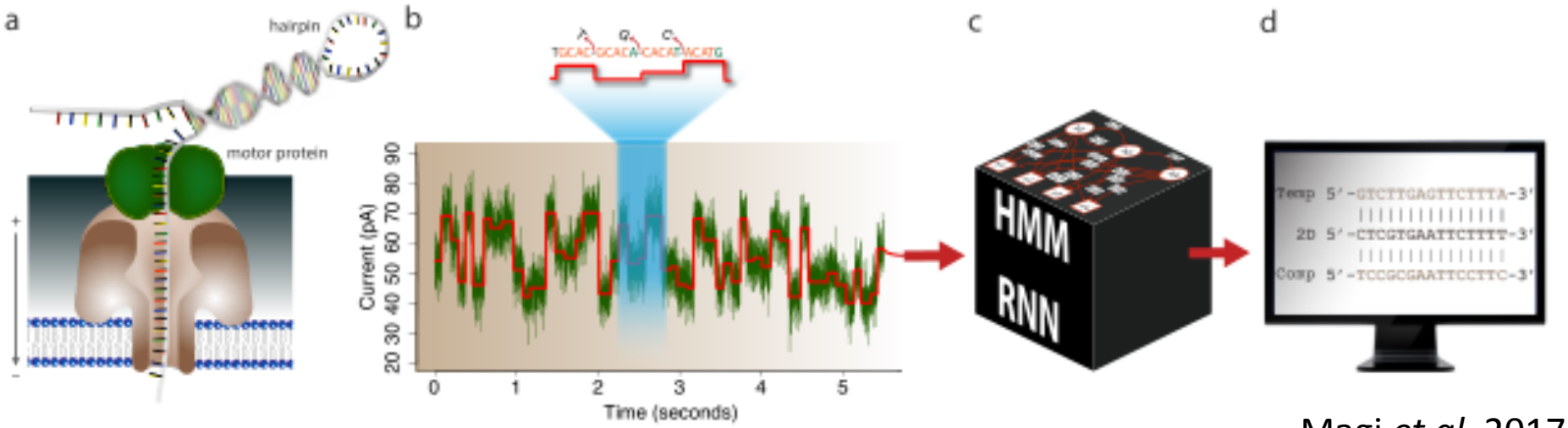
ion. ve been deposited in the NCBI r data are available at http:// tml.

psnyder@stanford.edu. pnas.org/lookup/suppl/doi:10.

11 | no. 27 | 9869–9874

Much better resolution of RNA sequences (e.g. splice variants)

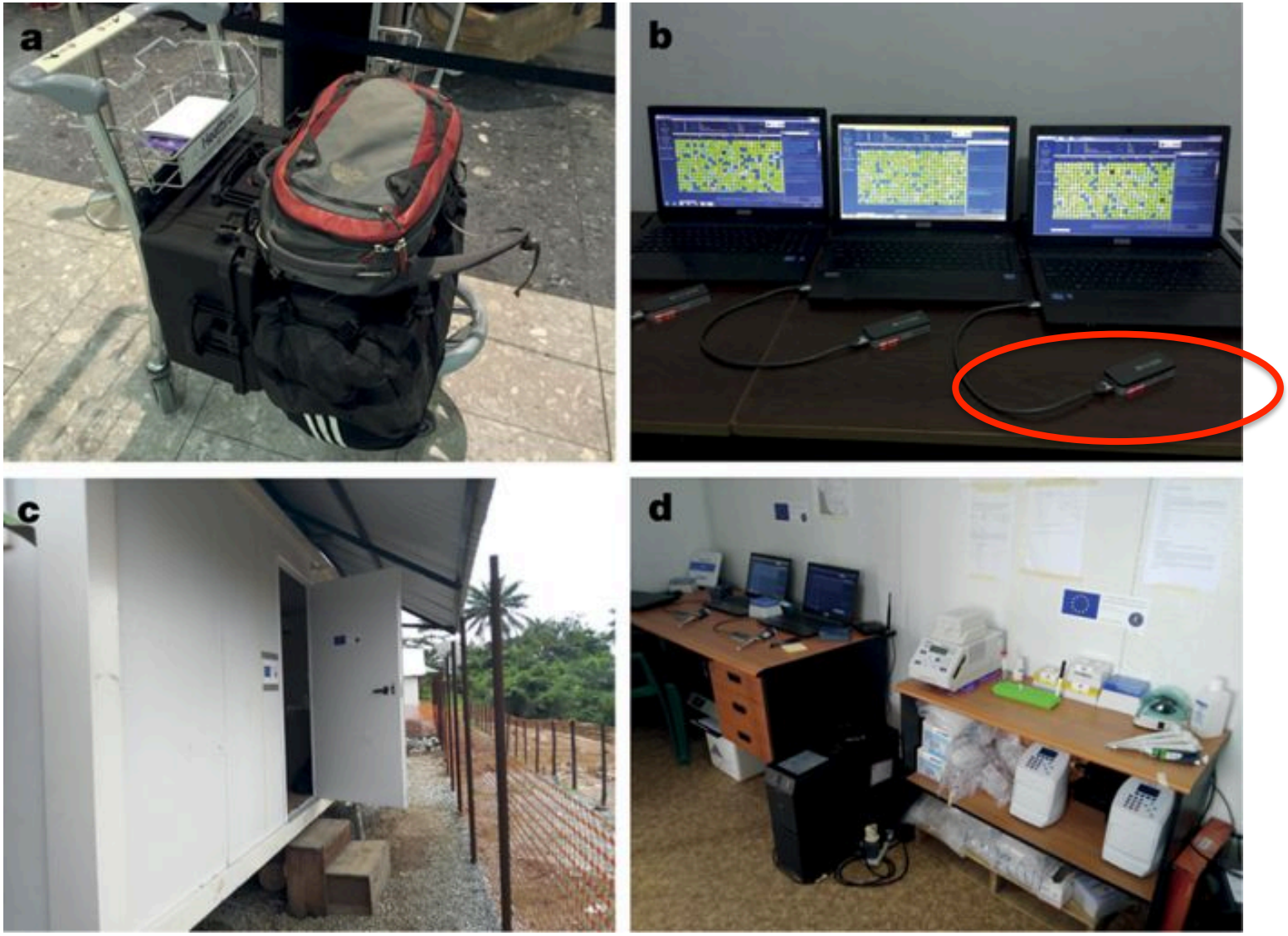
~2014: Nanopore sequencing



Magi *et al.* 2017

Not sequencing by synthesis: direct reading of the sequence

Nanopore sequencing



Quick *et al.* 2016

Using long-read sequencing to detect imprinted DNA methylation

Scott Gigante[1,a,e,✉], **Quentin Gouil**[1,a,b], **Alexis Lucattini**[c], **Andrew Keniry**[a], **Tamara Beck**[a], **Matthew Tinning**[c], **Lavinia Gordon**[c], **Chris Woodruff**[a], **Terence P. Speed**[a,d], **Marnie E. Blewitt**[2,a,b], **and Matthew E. Ritchie**[2,a,b,d,✉]

[1]Equal first author
[2]Equal last author
[a]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 3052 Australia
[b]Department of Medical Biology, The University of Melbourne, Parkville, 3010 Australia
[c]Australian Genome Research Facility, 305 Grattan Street, Melbourne, 3000 Australia
[d]School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010 Australia
[e]Department of Genetics, Yale University, New Haven, 06520 USA

Systematic variation in the methylation of cytosines at CpG sites plays a critical role in early development of humans and other mammals. Of particular interest are regions of differential methylation between parental alleles, as these often dictate monoallelic gene expression, resulting in parent of origin specific control of the embryonic transcriptome and subsequent development, in a phenomenon known as genomic imprinting.

Using long-read nanopore sequencing we show that, with an average genomic coverage of approximately ten, it is possible to determine both the level of methylation of CpG sites and the haplotype from which each read arises. The long-read property is exploited to characterise, using novel methods, both methylation and haplotype for reads that have reduced basecalling precision compared to Sanger sequencing. We validate the analysis both through comparison of nanopore-derived methylation patterns with those from Reduced Representation Bisulfite Sequencing data and through comparison with previously reported data.

Our analysis successfully identifies known imprinting control regions as well as some novel differentially methylated regions which, due to their proximity to hitherto unknown monoallelically expressed genes, may represent new imprinting control regions.

Nanopore sequencing | Differential methylation | Haplotyping | Imprinting | Long-read sequencing

Correspondence: ***scott.gigante@yale.edu, mritchie@wehi.edu.au***

Introduction

Methylation of the 5th carbon of cytosines (5mC or simply mC) is an epigenetic modification essential for normal mammalian development. Methylation differences between alleles contribute to establishing allele-specific expression patterns. As obtaining genome-wide haplotyped methylomes with short reads remains challenging, we evaluated the ability of long read, nanopore-based sequencing to improve allele-specific methylation analyses.

We apply the technique to the study of genomic imprinting, where differential expression of the maternal and paternal alleles in the offspring is at least partially set by the differential methylation (1–5). Imprinting is proposed to arise from the diverging interests of the maternal and paternal genes (6). In accordance with its primordial role in allocation of resources from the mother to the offspring, the placenta, along with the brain, is the organ where parental conflict results in the most pronounced imprinted expression (7–9). We thus conduct a survey of differential methylation and expression in murine embryonic placenta.

Recent studies have increased the number of genes identified as subject to imprinting in mouse to about 200 (10–15). The cause of the differential expression between paternal and maternal alleles is only known for a subset of these genes; maternal histone marks can play a role (14), and in other cases it involves the differential methylation of adjacent regions (5). The differential methylation patterns may be established in the gametes and persist through the epigenetic reprogramming occurring after fertilisation (16). These differentially methylated regions (DMRs) are called primary DMRs, or imprinting control regions (ICRs). Alternatively, differential methylation may arise during development, perhaps as a downstream effect of differential expression, in which case the regions are called somatic or secondary DMRs (17).

Apart from the parent of origin of the allele, genetic differences can also be associated with differential methylation. In this case, F1 hybrids of distinct mouse strains will display DMRs between the alleles according to the strain of origin (18), and not the parent. Genetically determined DMRs can have profound effects on phenotype, for instance in humans by altering the expression of mismatch repair genes important in cancer (19). Therefore, we also investigate the link between DNA methylation and expression for strain-biased genes.

Reconstructing haplotyped methylomes necessitates the simultaneous measurement of DNA methylation and single-nucleotide polymorphisms (SNPs) differentiating the alleles. This can be achieved by deep sequencing of bisulfite-converted DNA on the Illumina platforms, although the short reads combined with the reduced complexity of the bisulfite-treated DNA make the process inefficient, meaning many regions with low SNP density remain unresolved. Long reads provided by third generation sequencing technologies can overcome the requirement of a high SNP density, while several methods allow the assessment of base modifications on native DNA (thus also avoiding the reduction in complexity associated with bisulfite conversion). These methods include: analysis of polymerase kinetics for PacBio SMRT se-

DNA sequencing types: summary

Method	Read Length	Single pass error rate (%)	Reads per run	Time per run	Cost per million bases (USD)
Sanger (ABI)	600-1000	0.001	96	0.5 – 3h	500
454	700	1	1e6	23 h	8.57
Illumina (HiSeq)	2 x 125	0.1	8e9 (paired)	7 – 60 h	0.03
PacBio (RS II)	1-1.5e4	13	3.5-7.5e4	0.5 – 4 h	0.40-0.80
Oxford Nanopore (MinION)	2-5e3	38	1.1-4.7e4	50 h	6.44-17.90

Adapted from Rhoads & Au 2015

Data processing

Performed by sequencing team

Base calling (A/C/T/G)

Quality Scoring

Adapter trimming

FASTA/FASTQ files

Performed by researcher

Assembly

???

Alignment/mapping to reference

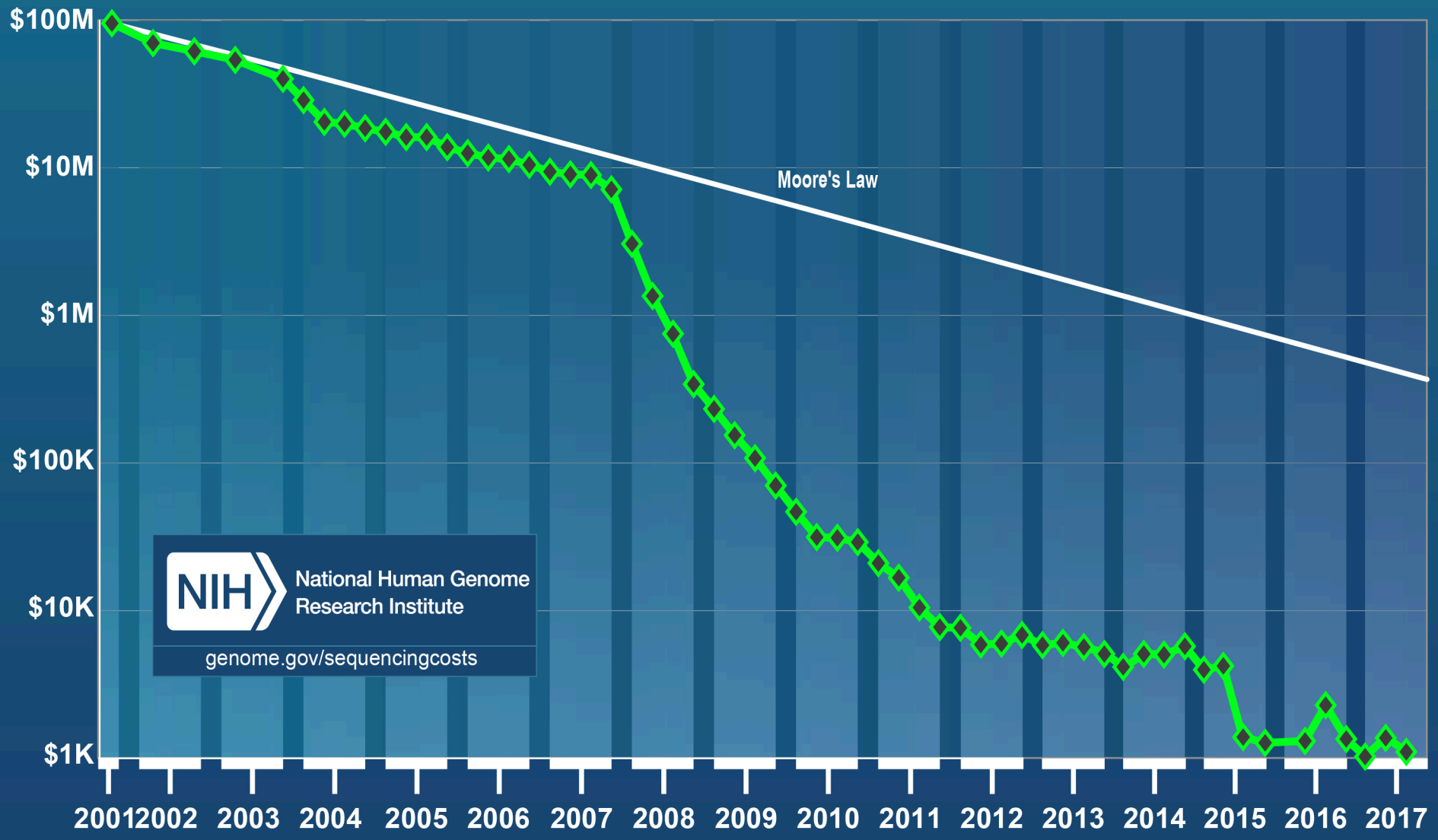
FASTA/FASTQ

```
@NB551431:31:HYTTGBGX7:1:23312:16016:20368 1:N:0:AGCGATAG+AGGCTATA
GGGGAAGTATGTAGGAGTTGAAGATTAGTCCGCCGTAGTCGGTGTACTCGTAGGTTCAGTACCATTGGTGGCCAAT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551431:31:HYTTGBGX7:1:23312:4412:20371 1:N:0:AGCGATAG+AGGCTATA
CCCCGTTCGGTCGGCACAGTTAGGACTCCCTCCCTGGGAGAGAATCACGACCCTGACTTAGAGGAAGACTCGACTC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/EEEEEE/EEEEEAEAEEEEEEEEEAEEEEEEEE
@NB551431:31:HYTTGBGX7:1:23312:5529:20373 1:N:0:AGCGATAG+AGGCTATA
GTTTTTGTTTTACTGCTGTGCTTGATATACATGAAGTAATGAATACCAAGCAATTCATTTTTCCTGCATCTTTACT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551431:31:HYTTGBGX7:1:23312:7028:20379 1:N:0:AGCGATAG+AGGCTATA
GCCTCCTTCTCAAATTTTTCAATGGTTCTTTTGTCGATGCCACCGCATTTATAGATCAGATGGCCAGTAGTGGTGG
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551431:31:HYTTGBGX7:1:23312:24086:20380 1:N:0:AGCGATAG+AGGCTATA
CTGGATTCCTGCACTGGCTGTGAACTTCTGCCAAGCTCCCCAGTCATCCTGGTCAAAGGGATCTTCGATAGACACC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551431:31:HYTTGBGX7:1:23312:22838:20386 1:N:0:NGCGATAG+AGGCTATA
ATTGGGTAAAAGATGAGCTAGCTGTTCTAGTATTTGCTTTTTGTAATCCAGTTAAGACCATGAGCATATACAATAT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Current challenges

- Repeats, especially homopolymers, are still difficult to sequence
  - Lower quality scores, less coverage
- Read lengths are still significantly shorter than some RNA transcripts
- Most technologies require amplification before sequencing, which can introduce errors and biases
- Most technologies don't provide information about base modifications
- Some technologies require large amounts of input material
- Cost, time, convenience

Cost per Genome



https://www.genome.gov/sequencingcostsdata/

References

- Avery OT *et al.* 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation desoxyribonucleic acid fraction isolated from pneumococcus type III. Journal of Experimental Medicine 79, 137–157
- Dahm R 2005. Freidrich Miescher and the discovery of DNA. Developmental Biology 278(2):274-288
- Griffith F. 1928. The significance of pneumococcal types. Journal of Hygiene 27: 113-159
- Johnson DS *et al.* Genome-wide mapping of in vivo protein-DNA Interactions. Science 316(5830): 1497-1502
- MacDonald ME *et al.* 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72(6): 971-983.
- Magi A *et al.* 2017. Nanopore sequencing data analysis: state of the art, applications and challenges. Briefings in Bioinformatics.
- Margulies M *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors Nature 437: 376-380
- Pasini EA *et al.* 2017. An improved Plasmodium cynomolgi genome assembly reveals an unexpected methyltransferase gene expansion. Wellcome Open Research 2:42.
- Quick J *et al*. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530: 228-232
- Ratner L 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313: 277-284.
- Rhoads A and Au KF. 2015. PacBio sequencing and its applications. Genomics, Proteomics and Bioinformatics 13(5):278-289
- Riordan JR *et al.* 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245(4922): 1066-1073.
- Sanger F *et al.* 1977. DNA sequencing with chain-terminating inhibitors. PNAS 74(12): 5463-5467
- Sanger F *et al.* 1977. Nucleotide sequence of bacteriophage ϕX174. Nature 265: 687-695
- Tilgner H *et al.* 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. PNAS 111(27): 9869-9874.
- Valln HKM and Caldas C. 2011. The breast cancer genome – a key for better oncology. BMC Cancer 11: 501.
- Watson JD and Crick FHC. 1953. A structure for deoxyribose nucleic acid. Nature 171: 737–738