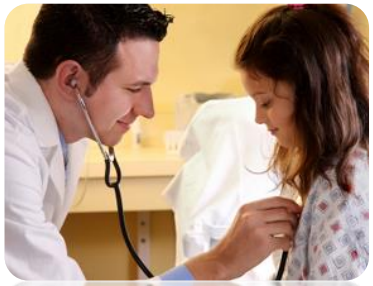


長短兼備：
高通量DNA定序技術於生物醫學研究之應用
Short and Long Read Sequencing Technologies and their Applications in Biomedical Research

黃憲達 (Hsien-Da Huang)

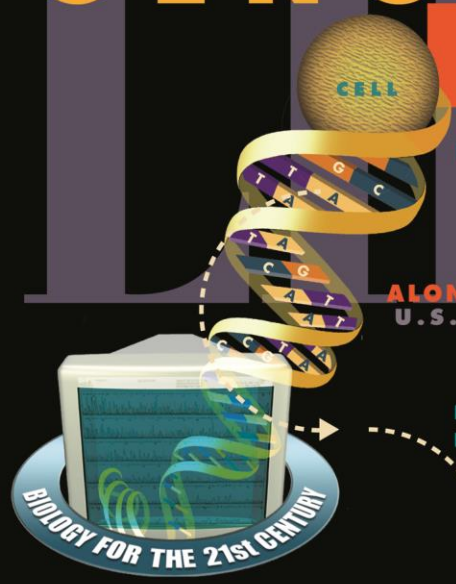
國立交通大學 生物科技學系 講座教授
國立交通大學 生物科技學院 副院長

GENOMES *to* LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

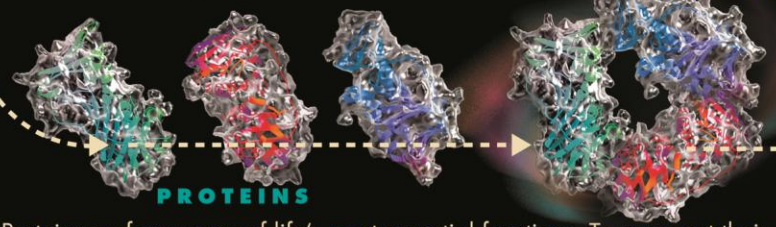
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

U.S. DEPARTMENT OF ENERGY



DNA SEQUENCE DATA FROM GENOME PROJECTS

Genes and other DNA sequences contain instructions on how and when to build proteins

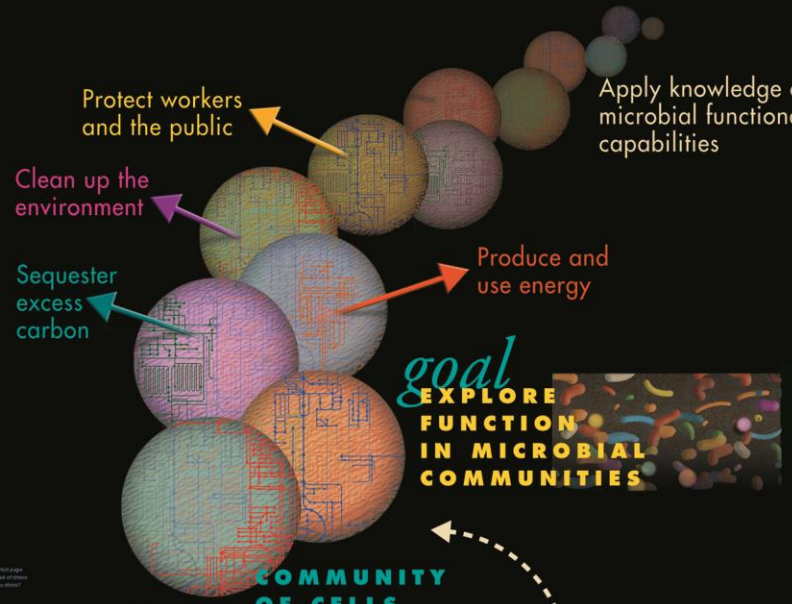


PROTEINS

Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

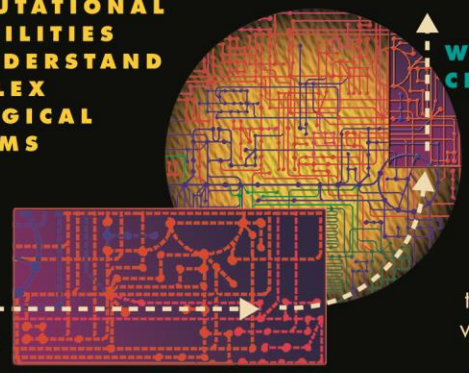
goal IDENTIFY PROTEIN MACHINES

goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS



COMMUNITY OF CELLS

goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES



WORKING CELL

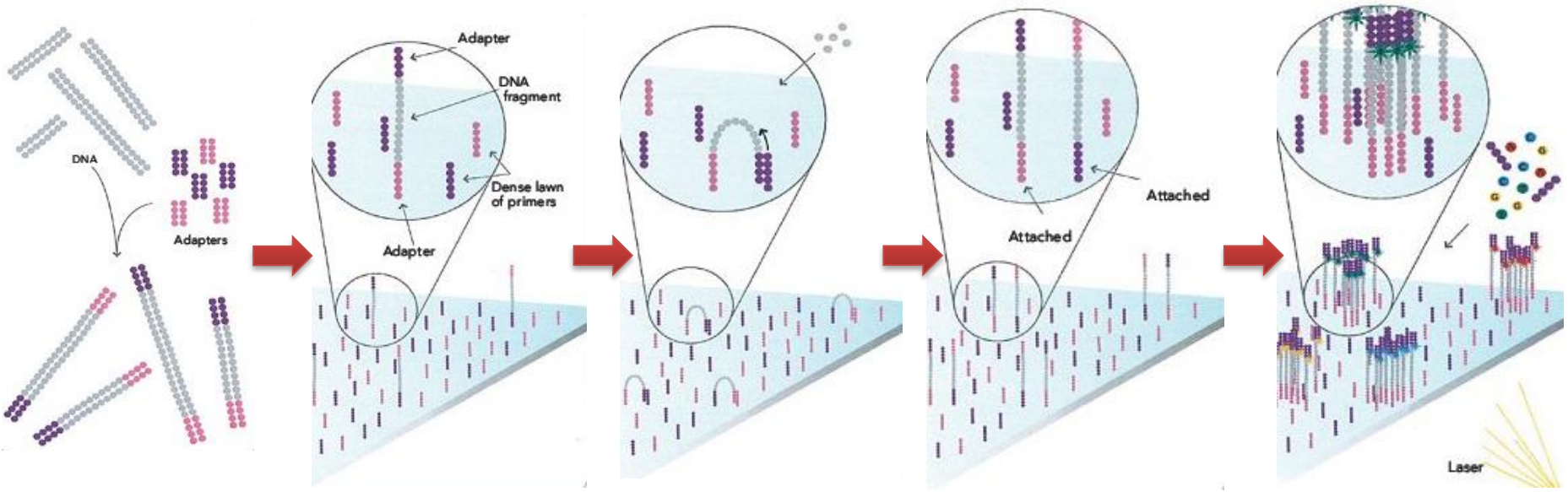
Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.

goal CHARACTERIZE GENE REGULATORY NETWORKS

URL *DOEGenomesToLife.org*

10/02

Next-generation sequencing



GCGCAGCGCGGCCGCAGCAGCCTCCGCCCCCCGCACGGTGTC

GTCCCGAGCTAGCCCCGGCGGCCGCCGCCGCCCAGACCGGAC

AGTCCCCGCCTCGCCGCCAACGCCACAACCACCGCGCACGGC

AGAGCCGGAGCGAGCTCTTCGGGGAGCAGCGATGCGACCCTC

CTGCTGGCTGCGCTCTGCCCGGCGAGTCGGGCTCTGGAGGAA

ACGGCTCGTGCGTCCGAGCCTGTGGGGCCGACAGCTATGAGA

GAAGTGCGAAGGGCCTTGCCGCAAAGTGTGTAACGGAATAGC

ATAAATGCTACGAATATTAAACACTTCAAAAACTGCACCTCC

TGGCATTTAGGGGTGACTCCTTCACACATACTCCTCCTCTGC

CGTAAAGGAAATCACAGGGTTTTTGCTGATTCAGGCTTGGCC

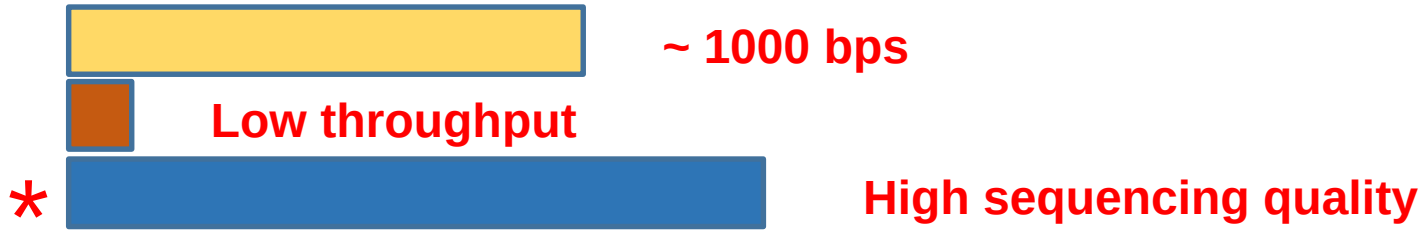
GAGAACCTAGAAATCATACGCGGCAGGACCAAGCAACATGGT

ACATAACATCCTTGGGATTACGCTCCCTCAAGGAGATAAGTC

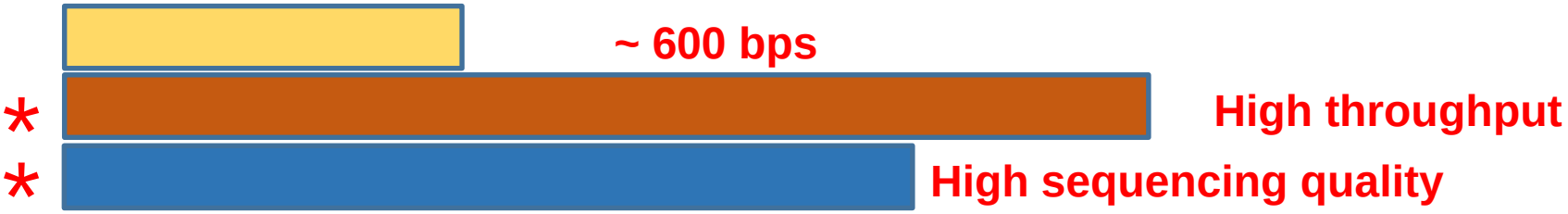
AAATTTGTGCTATGCAAATACAATAAACTGGAAAAAACTGTT

Sequencing Length, Throughput & Quality

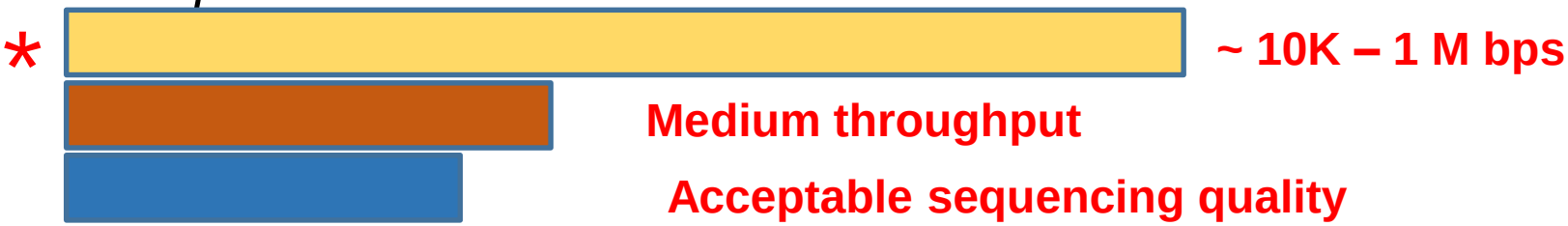
Sanger sequencing



Next-generation sequencing (NGS): illumina platform, Thermo Fisher



3rd or 4th generation sequencing: Pac Bio Science, Oxford Nanopore



豪傑使長槍、君王用短劍

「十八般武器」是傳統兵器總稱。各有形製、功能不同，如「刀」下就有九環刀、雁翎刀、青龍偃月刀等。

武器使用的訣竅，依不同功能而設計不同，如「槍」用於遠刺；「劍」以刺、割為主。

「十八般武器」分成「九長、九短」，「九長」是：刀、槍、棍、鉞、叉、鑭、鉤、槊、戟。「九短」是：刀、劍、鞭、槁、枴、斧、棒、鎚、杵。

曹操-青釭劍

楊鐵心-耍長槍

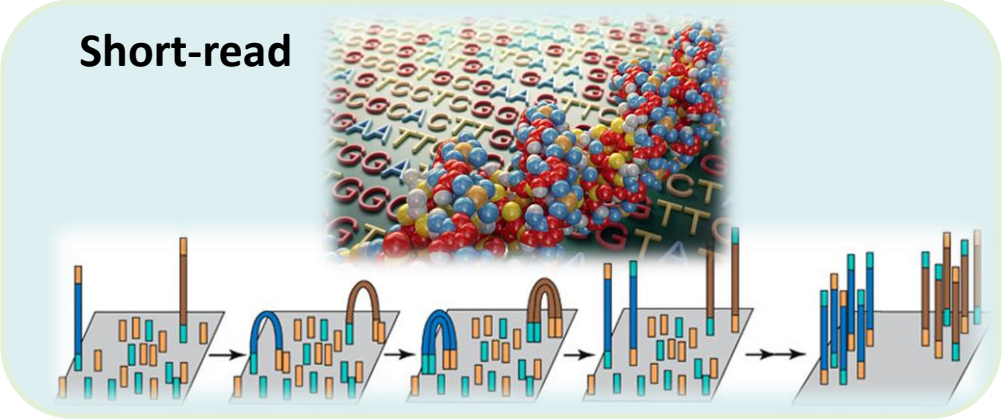
七十二路「楊家槍法」，出槍長，且虛實，有奇正，進其銳，退其速，其勢險，其節短，穩如山，動如雷。

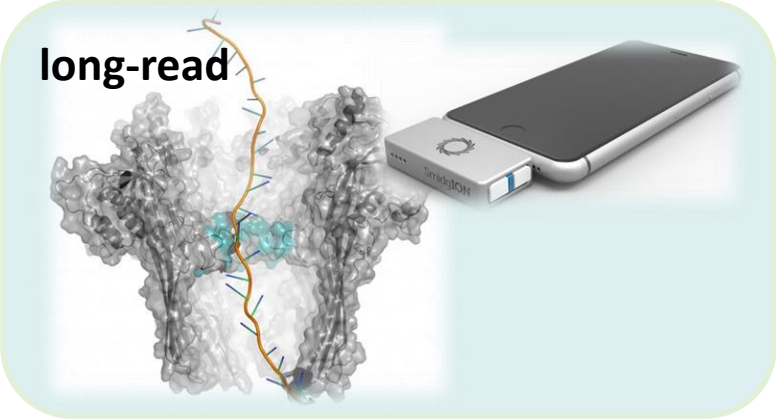
劉備-雙股劍

孫權-青冥劍

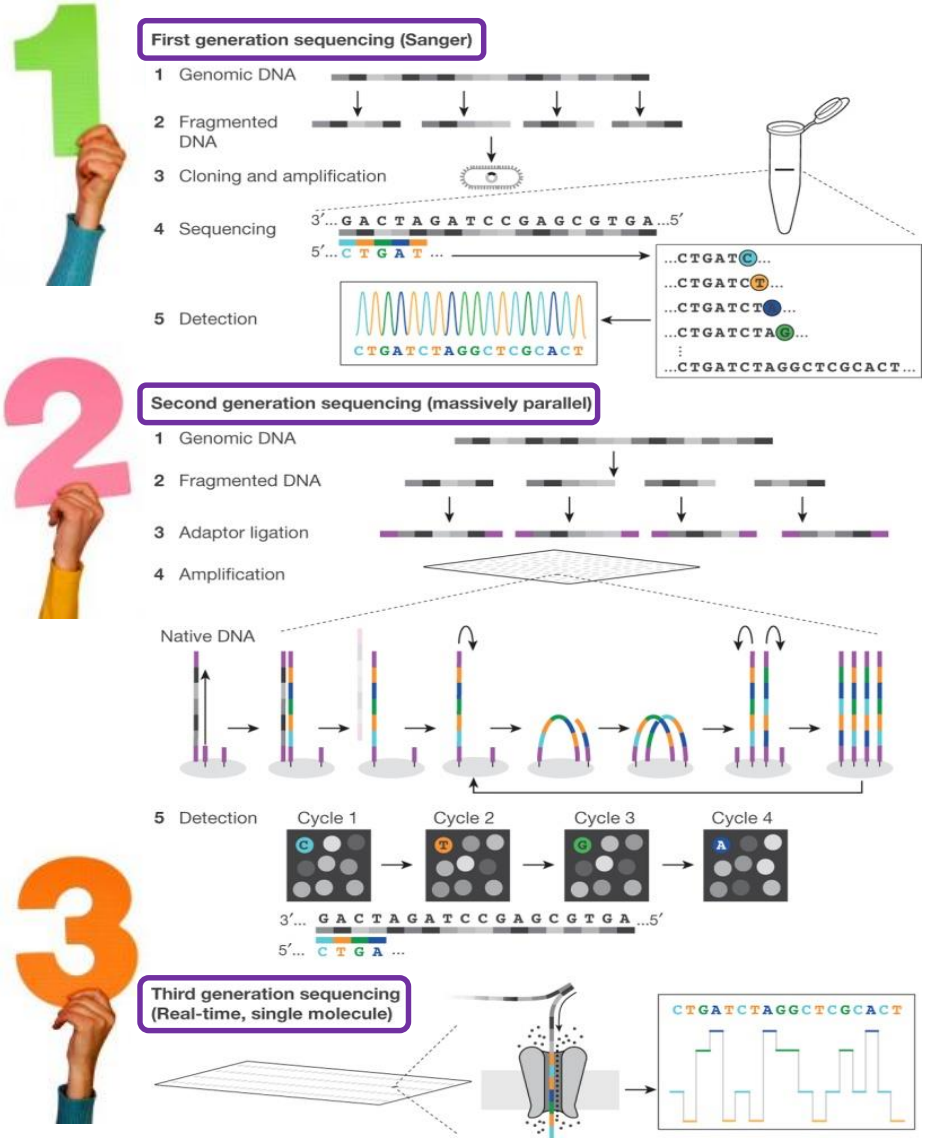
Short-read vs. long-read sequencing
Which one is better?

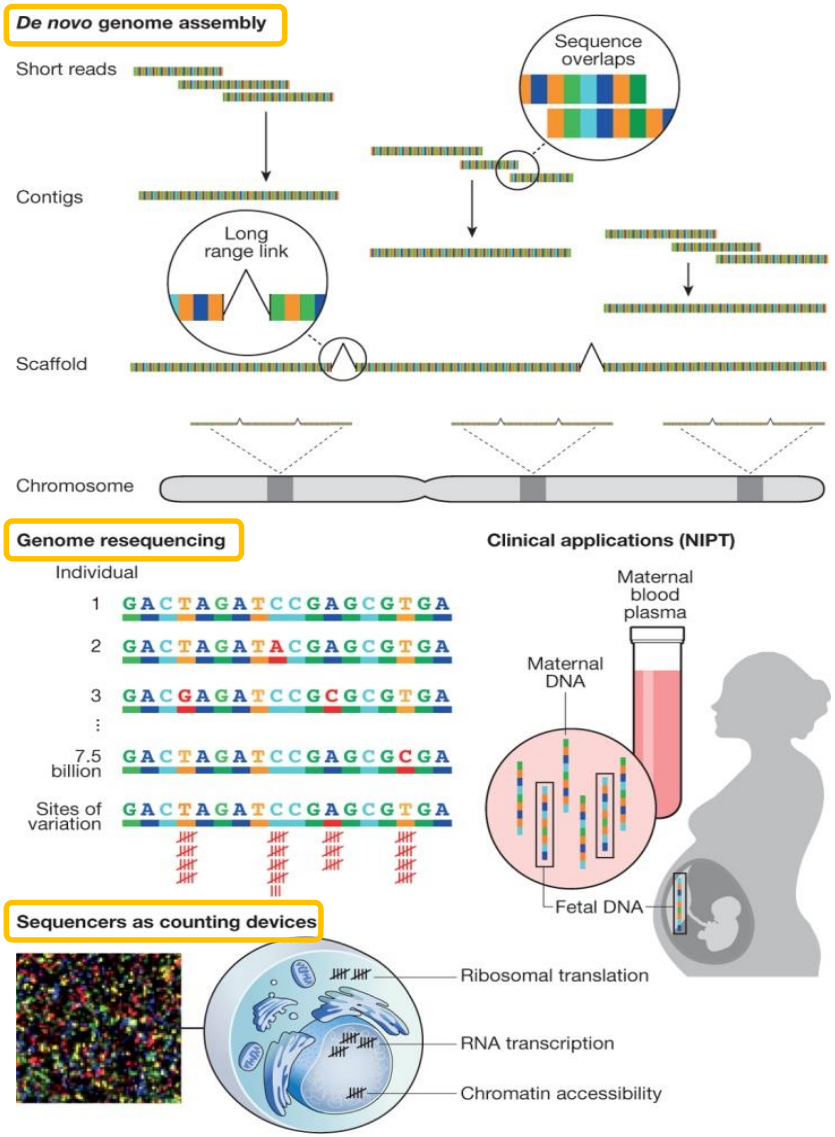
- The **most frequently asked questions** in sequencing.
- The sequencing read length depends on the instrument and chemistry used.
- The range of the read length of a short-read sequencing instrument is between 100 and 600 bps, while that of a long-read sequencing instrument varies between 10 to 15 kbps.
- The choice you make depends on the **goal of your experiment**; one isn't considered universally superior to the other.





DNA sequencing technologies and applications





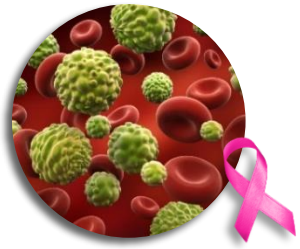
Nature 1–9 (2017) doi:10.1038/nature24286

Applications of Next-generation Sequencing

Human Genetics Research

Forensic Genomics

Complex
Disease Genomics



Oncology

Using NGS Technology

Reproductive Health

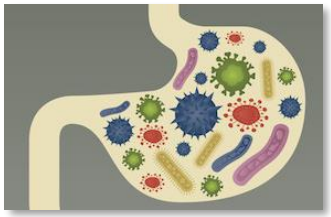
Genomics in
Drug Development

Agrigenomics Research

Agrigenomics

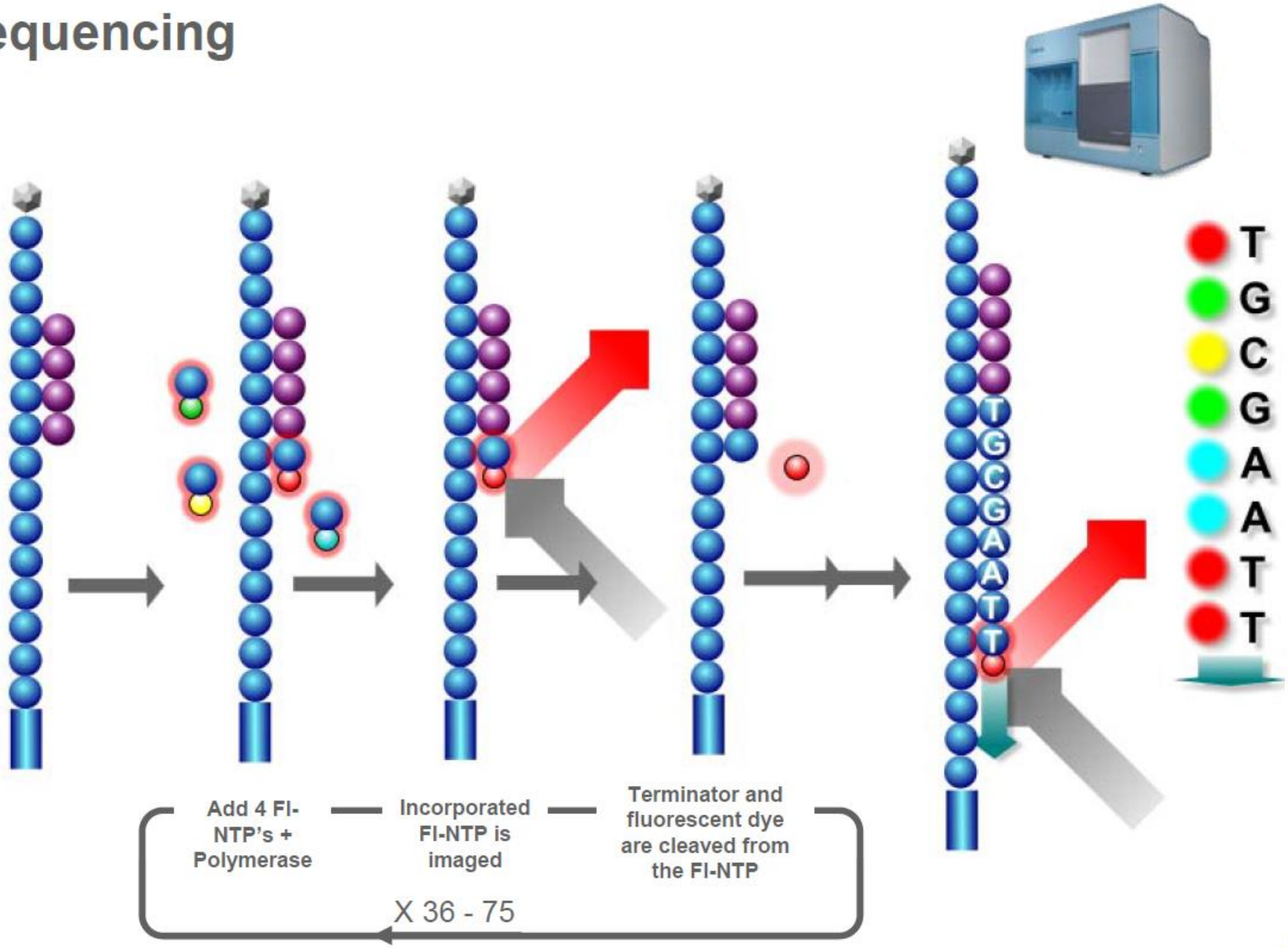
Metagenomics Research

Microbial Genomics



Microbiota

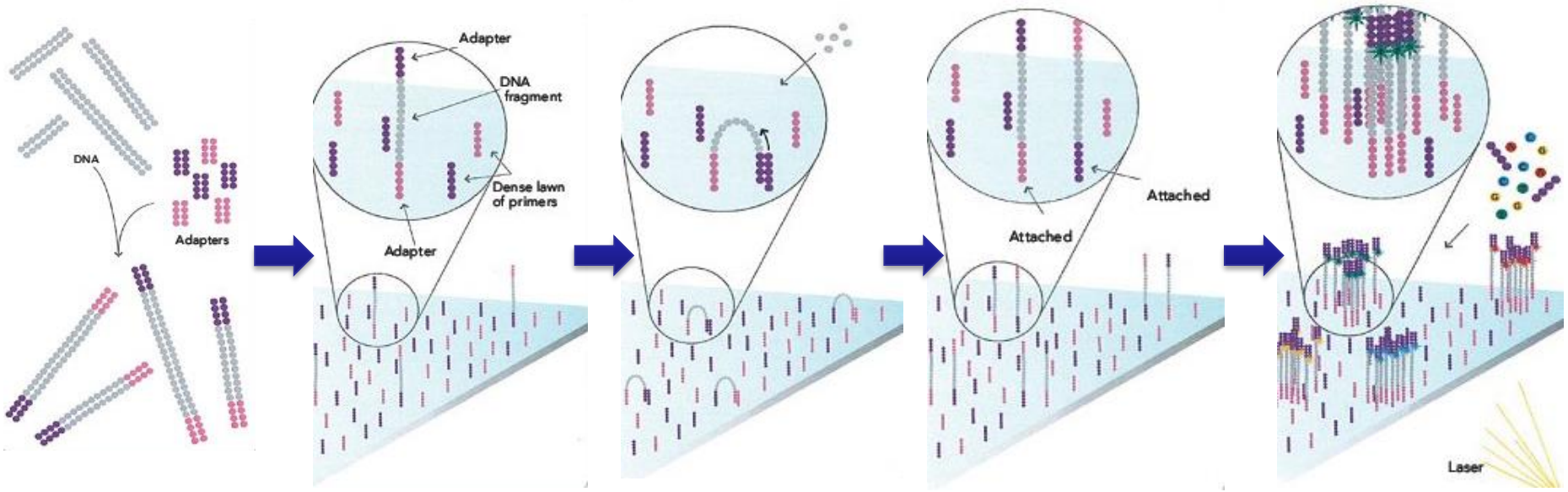
Sequencing



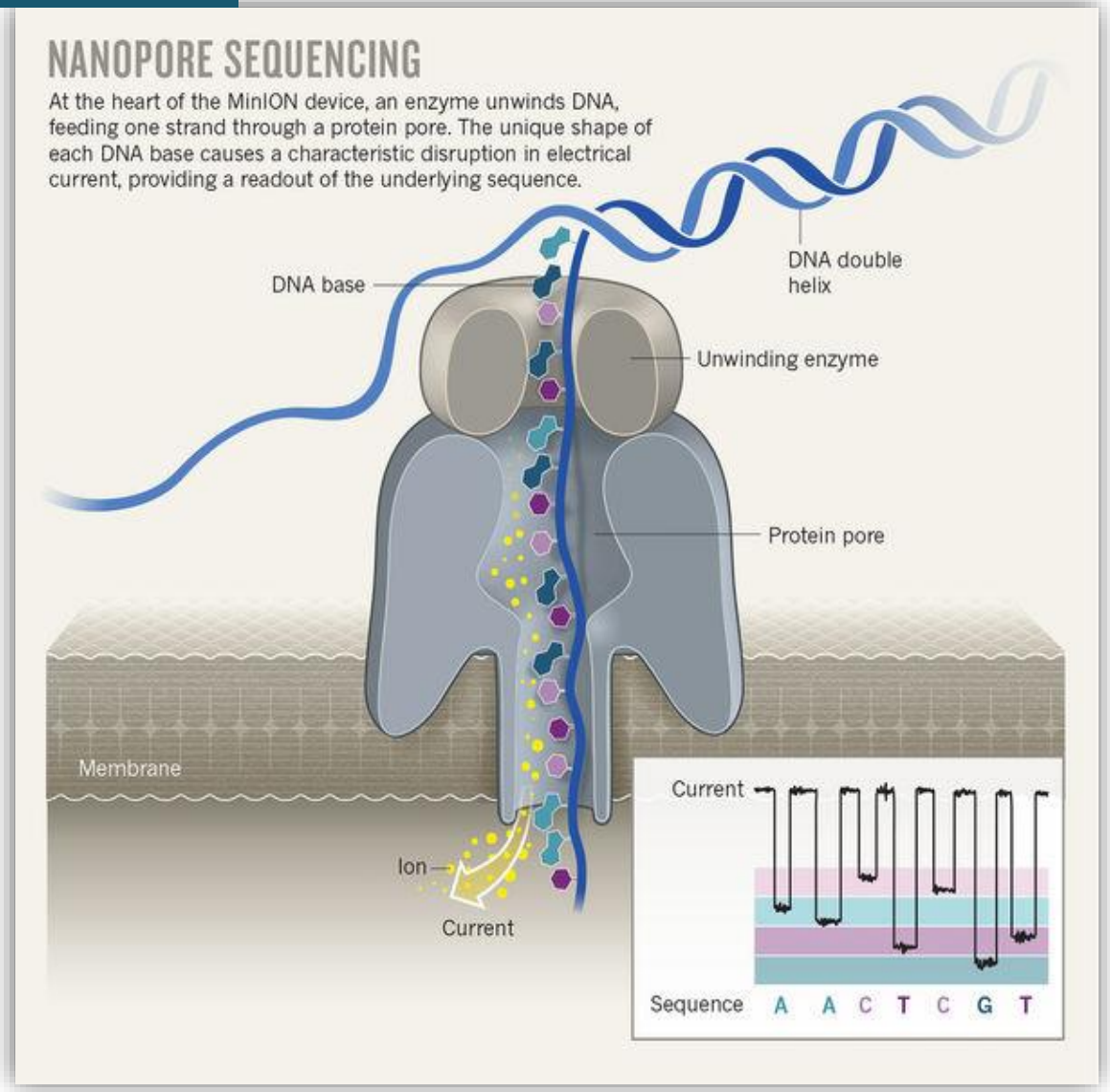
illumina

http://mi.caspur.it/workshop_NGS09/docs/Cappelletti_NGS09.pdf

Next-generation sequencing (NGS)

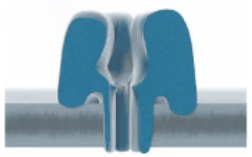


From **www.illumina.com**



TechBlog: The nanopore toolbox. 16 Oct 2017, Posted by Jeffrey Perkel

Oxford Nanopore Technology (ONT)



- 超長讀長，取長補短
- 隨測隨停，通量靈活
- 鹼基修飾，實時讀取
- RNA分子，直接測序

https://nanoporetech.com/resource-centre/videos/minion-introduction

Assessing the quality of a Sequencing Read

Phred quality scores

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

$$Q = -10 \log_{10}(P)$$

Q: Phred quality score
P: Base Call error rate

Genome Res. March 1, 1998 8: 186-194; doi:10.1101/gr.8.3.186

Read of fastq format

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Phred Q score

http://en.wikipedia.org/wiki/FASTQ_format

Sequencing Platform Comparison

Platform Comparison	Illumina	Thermo Fisher	Pacific Biosciences	Oxford Nanopore
Sequencing by Synthesis	Yes	Yes	Yes	No
DNA Size Selection	Yes	Yes	Yes	No
Post-library Amplification	Yes	Yes	No, Single Molecule	No, Single Molecule
Detection	Fluorescent Imaging	Ion Semiconductor	Fluorescent Imaging	Ionic Current Change
Sequencing Rate (s/base)	2 – 20 sec	30 sec	0.25 sec	0.002 sec
Running Time	Fixed	Fixed	Fixed	Run & Stop
DNA Sequencing	Yes	Yes	Yes	Yes
Direct DNA Modification Detection	No	No	Yes	Yes
Direct RNA Sequencing	No	No	No	Yes
Read Length	Short, up to 300 bp X 2	Short, up to 600 bp	Long, Average 6-8 Kb	Long, Average 6-30 Kb
Total Reads (M)	4 – 800 (PE)	2 – 130	0.3 – 0.5	0.3 – 0.5
Total Base (Gb)	1.2 - 120	0.3 – 25 /Chip	5 – 8 /SMRT Cell	2 – 10 /Flow Cell
Instrument Cost (USD)	20K – 275K	200 – 300K	350K	1K – 125K

What Short Read NGS Cannot Do?

LOW DIAGNOSTIC YIELD OF CURRENT STATE-OF-THE-ART NGS BASED TEST

Jamuar and Tan *Human Genomics* (2015) 9:10
DOI 10.1186/s40246-015-0031-5

25%

Human Genomics

REVIEW — Open Access

Clinical application of next-generation sequencing for Mendelian diseases

Saumya Shekhar Jamuar[1,2] and Ene-Choo Tan[2,3*]

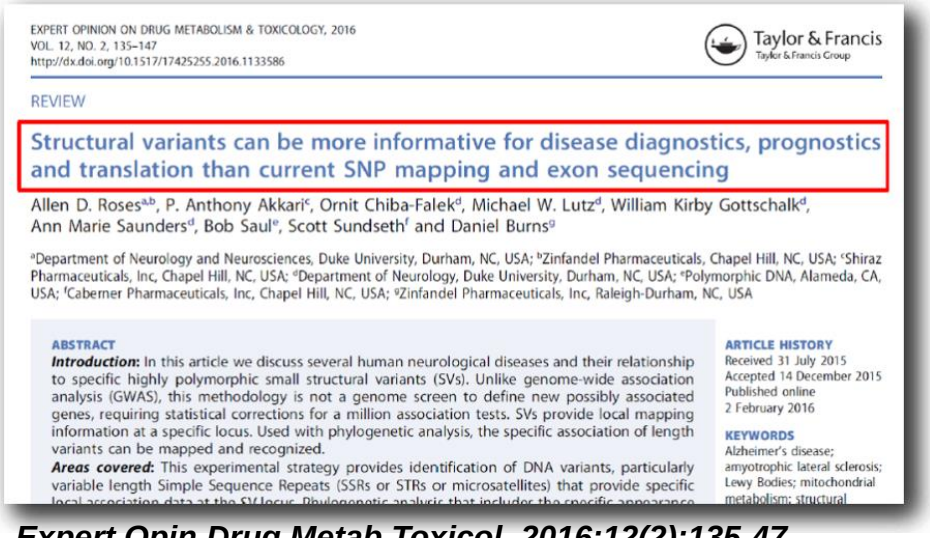
Abstract

Over the past decade, next-generation sequencing (NGS) has led to an exponential increase in our understanding of the genetic basis of Mendelian diseases. NGS allows for the analysis of multiple regions of the genome in one single reaction and has been shown to be a cost-effective and efficient tool in investigating patients with Mendelian diseases. More recently, NGS has been successfully deployed in the clinics, with a reported diagnostic yield of ~25 %. However, recommendations on clinical implementation of NGS are still evolving with numerous key challenges that impede the widespread use of genetics in everyday medicine. These challenges include when to order, on whom to order, what type of test to order, and how to interpret and communicate the results, including incidental findings, to the patient and family. In this review, we discuss these challenges and suggest guidelines on implementing NGS in the routine clinical workflow.

Diagnostic yield: The likelihood that a test or procedure will provide the information needed to establish a diagnosis.

Hum Genomics. 2015; 9(1): 10.

STRUCTURAL VARIATION MORE IMPORTANT THAN SNP



EXPERT OPINION ON DRUG METABOLISM & TOXICOLOGY, 2016
VOL. 12, NO. 2, 135–147
http://dx.doi.org/10.1517/17425255.2016.1133586

Taylor & Francis
Taylor & Francis Group

REVIEW

Structural variants can be more informative for disease diagnostics, prognostics and translation than current SNP mapping and exon sequencing

Allen D. Roses[a,b], P. Anthony Akkari[c], Ornit Chiba-Falek[d], Michael W. Lutz[d], William Kirby Gottschalk[d], Ann Marie Saunders[d], Bob Saul[e], Scott Sundseth[f] and Daniel Burns[g]

[a]Department of Neurology and Neurosciences, Duke University, Durham, NC, USA; [b]Zinfandel Pharmaceuticals, Chapel Hill, NC, USA; [c]Shiraz Pharmaceuticals, Inc, Chapel Hill, NC, USA; [d]Department of Neurology, Duke University, Durham, NC, USA; [e]Polymorphic DNA, Alameda, CA, USA; [f]Cabernet Pharmaceuticals, Inc, Chapel Hill, NC, USA; [g]Zinfandel Pharmaceuticals, Inc, Raleigh-Durham, NC, USA

ABSTRACT

Introduction: In this article we discuss several human neurological diseases and their relationship to specific highly polymorphic small structural variants (SVs). Unlike genome-wide association analysis (GWAS), this methodology is not a genome screen to define new possibly associated genes, requiring statistical corrections for a million association tests. SVs provide local mapping information at a specific locus. Used with phylogenetic analysis, the specific association of length variants can be mapped and recognized.

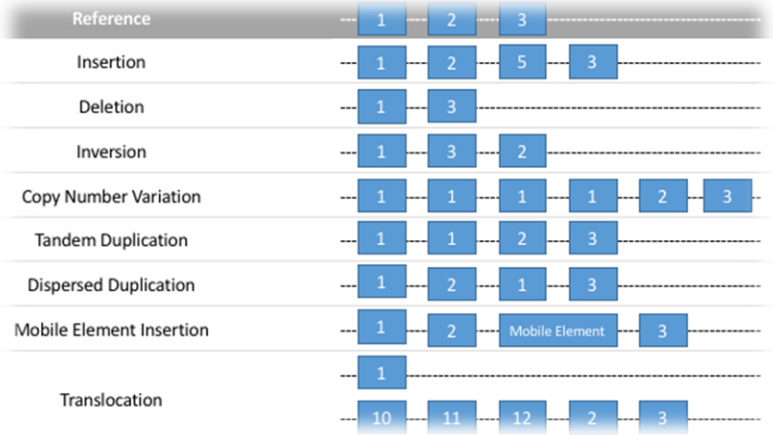
Areas covered: This experimental strategy provides identification of DNA variants, particularly variable length Simple Sequence Repeats (SSRs or STRs or microsatellites) that provide specific

ARTICLE HISTORY
Received 31 July 2015
Accepted 14 December 2015
Published online
2 February 2016

KEYWORDS
Alzheimer's disease; amyotrophic lateral sclerosis; Lewy Bodies; mitochondrial metabolism; structural

Expert Opin Drug Metab Toxicol. 2016;12(2):135-47.

complex structural variations



https://en.wikipedia.org/wiki/Human_Genome_Structural_Variation

Structural Variants are More Impactful than SNP

EXPERT OPINION ON DRUG METABOLISM & TOXICOLOGY, 2016
VOL. 12, NO. 2, 135–147
http://dx.doi.org/10.1517/17425255.2016.1133586

Taylor & Francis
Taylor & Francis Group

REVIEW

Structural variants can be more informative for disease diagnostics, prognostics and translation than current SNP mapping and exon sequencing

Allen D. Roses[a,b], P. Anthony Akkari[c], Ornit C
Ann Marie Saunders[d], Bob Saul[e], Scott Sund

[a]Department of Neurology and Neurosciences, Duke Uni
Pharmaceuticals, Inc, Chapel Hill, NC, USA; [d]Department
USA; [f]Caberner Pharmaceuticals, Inc, Chapel Hill, NC, US

ABSTRACT

Introduction: In this article we discuss several hu
to specific highly polymorphic small structural
analysis (GWAS), this methodology is not a ger
genes, requiring statistical corrections for a milli
information at a specific locus. Used with phylog
variants can be mapped and recognized.

Areas covered: This experimental strategy prov
variable length Simple Sequence Repeats (SSRs
local association data at the SV locus. Phylogeno

Expert Opin Drug Metab Toxicol. 2016;12(2):135-47.

ORIGINS OF REDUCED ACCURACY IN CLINICAL GENOMICS FROM SHORT SEQUENCING READS

nature REVIEWS GENETICS

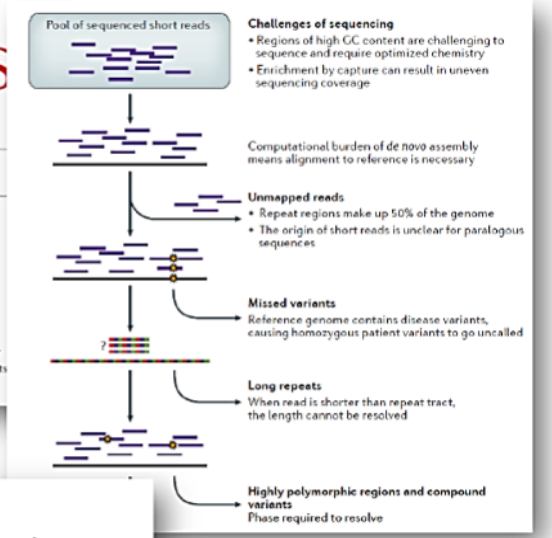
REVIEWS

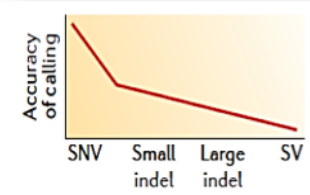
APPLICATIONS OF NEXT-GENERATION SEQUENCING

Towards precision medicine

Euan A. Ashley

Abstract | There is great potential for genome sequencing to enhance patient care through improved diagnostic sensitivity and more precise therapeutic targeting. To maximize this potential, genomics strategies that have been developed for genetic discovery — including DNA-sequencing technologies and analysis algorithms — need to be adapted to fit clinical needs. This will require the optimization of alignment algorithms, attention to quality-coverage metrics, tailored solutions for paralogous or low-complexity areas of the genome, and the adoption of consensus standards for variant calling and interpretation. Global sharing of this more accurate genotypic and phenotypic data will accelerate the determination of causality for novel genes or variants. Thus, a deeper understanding of disease will be realized that will allow its targeting with much greater therapeutic precision.





Accuracy of variant calling falls with increasing disruption of the open reading frame

Nat Rev Genet. 2016 Aug 16;17(9):507-22.

We Need a Deeper Genome

LOW *DIAGNOSTIC YIELD OF CURRENT STATE-OF-THE-ART NGS BASED TEST

Jamuar and Tan *Human Genomics* (2015) 9:10
DOI 10.1186/s40246-015-0031-5

Human Genomics

REVIEW **Open Access**

Clinical application of next-generation sequencing for Mendelian diseases

Saumya Shekhar Jamuar[1,2] and Ene-Choo Tan[2,3*]

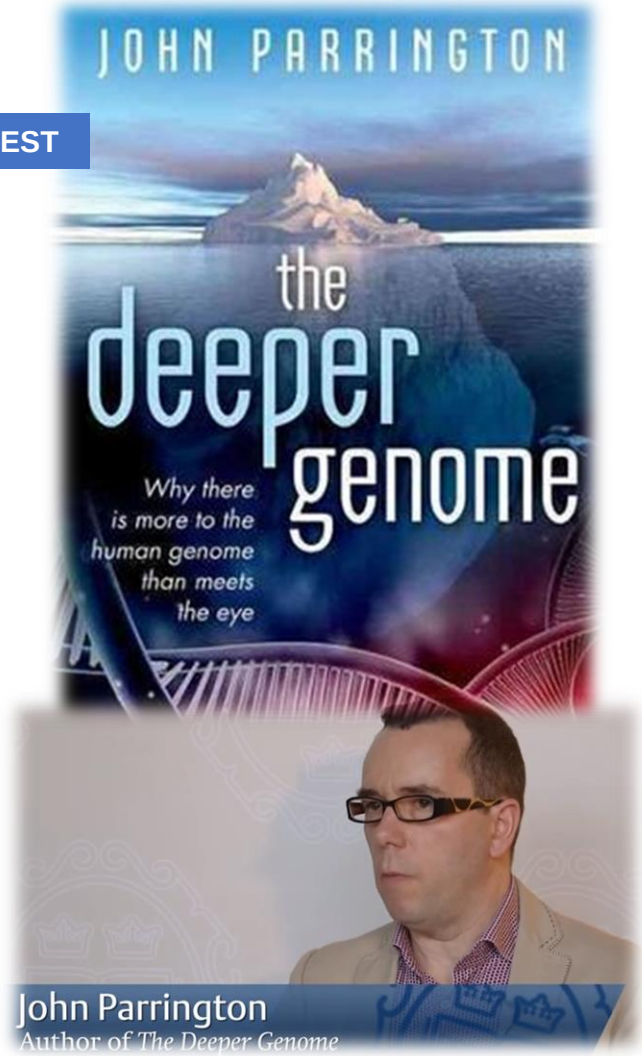
Abstract

Over the past decade, next-generation sequencing (NGS) has led to an exponential increase in our understanding of the genetic basis of Mendelian diseases. NGS allows for the analysis of multiple regions of the genome in one single reaction and has been shown to be a cost-effective and efficient tool in investigating patients with Mendelian diseases. More recently, NGS has been successfully deployed in the clinics, with a reported diagnostic yield of ~25 %. However, recommendations on clinical implementation of NGS are still evolving with numerous key challenges that impede the widespread use of genetics in everyday medicine. These challenges include when to order, on whom to order, what type of test to order, and how to interpret and communicate the results, including incidental findings, to the patient and family. In this review, we discuss these challenges and suggest guidelines on implementing NGS in the routine clinical workflow.

Keywords: Next-generation sequencing, Whole exome sequencing, Clinical applications, Mendelian diseases

***Diagnostic Yield:** the likelihood that a test or procedure will provide the information needed to establish a diagnosis

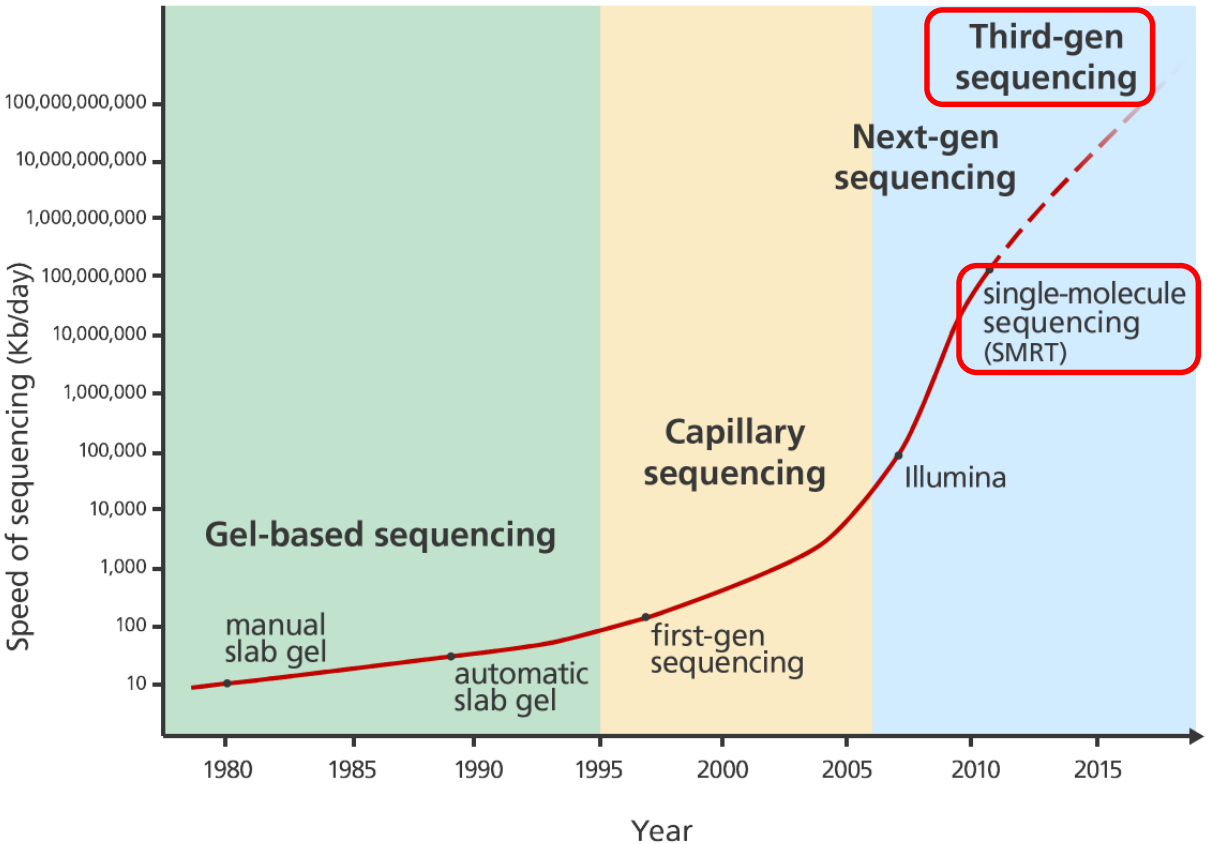
Hum Genomics. 2015; 9(1): 10.



https://www.youtube.com/watch?v=WYvuAtQo-_g

What's Beyond Short Read NGS?

HUMAN GENOMIC REGIONS UNRESOLVED BY SHORT SEQUENCING READS



https://www.yourgenome.org/stories/third-generation-sequencing

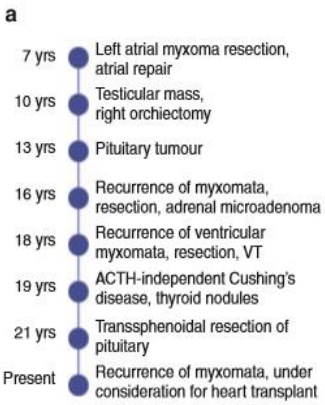
- Repeated regions (simple repeats, tandem repeats, transposon-related repeats)
- Highly polymorphic regions (HLA) – haplotype phasing
- Structural variants (relocation, inversion, duplications)
- Un-sequencable regions by NGS (AT or GC rich regions)

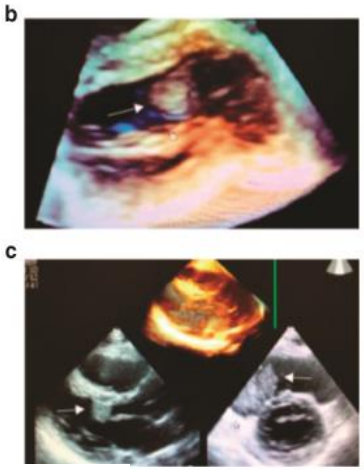
How “Long Reads” can Help?

BRIEF REPORT | Genetics inMedicine

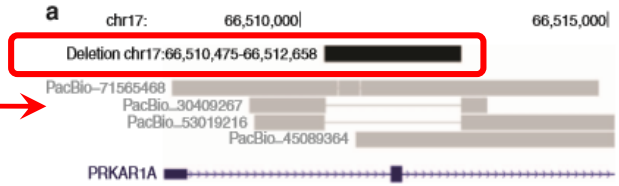
Long-read genome sequencing identifies causal structural variation in a Mendelian disease

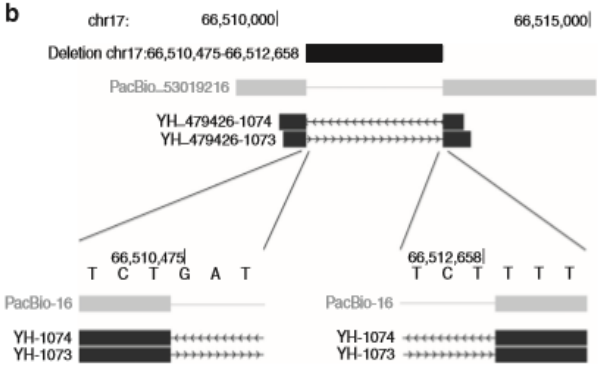
Jason D. Merker, MD, PhD[1,2], Aaron M. Wenger, PhD[3], Tam Sneddon, DPhil[2], Megan Grove, MS, LCGC[2], Zachary Zappala, PhD[1,4], Laure Fresard, PhD[1], Daryl Waggott, MSc[5,6], Sowmi Utiramerur, MS[2], Yanli Hou, PhD[1], Kevin S. Smith, PhD[1], Stephen B. Montgomery, PhD[1,4], Matthew Wheeler, MD, PhD[5,6], Jillian G. Buchan, PhD[1,2], Christine C. Lambert, BA[3], Kevin S. Eng, MS[3], Luke Hickey, BS[3], Jonas Korlach, PhD[3], James Ford, MD[4,5,7] and Euan A. Ashley, MRCP, DPhil[2,4,5,6]

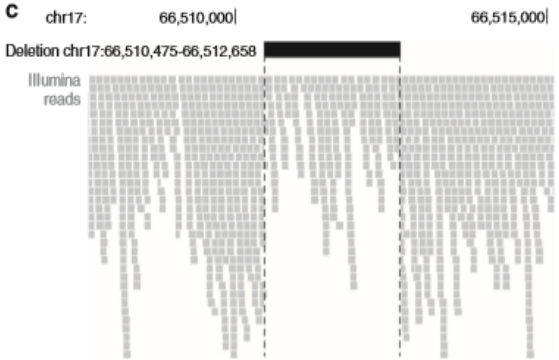




heterozygous 2,184 bp deletion →



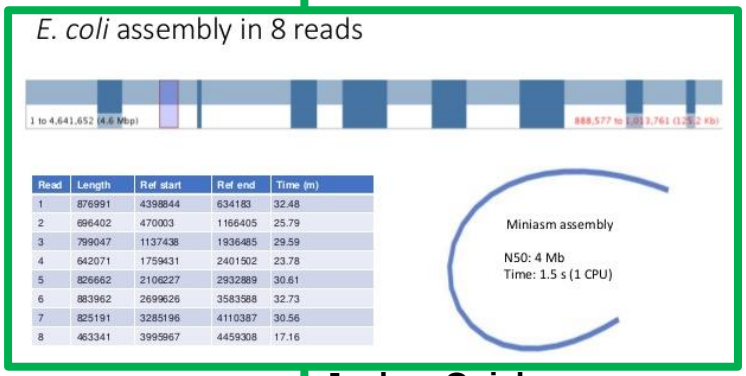




Genet Med. 2018 Jan;20(1):159-163.

Features and Limitations of Nanopore Sequencing

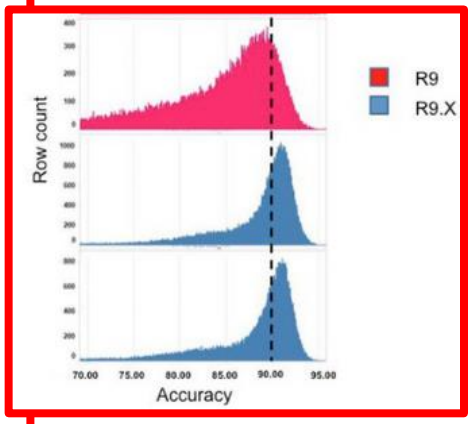
- Very long read-length
- Low capital investment
- Single molecule, PCR-free
- Real-time & urgent applications



Read	Length	Ref start	Ref end	Time (m)
1	876991	4398844	634183	32.48
2	696402	470003	1166405	25.79
3	799047	1137438	1936485	29.59
4	642071	1759431	2401502	23.78
5	826662	2106227	2932889	30.61
6	883962	2699626	3583588	32.73
7	825191	3285196	4110387	30.56
8	463341	3995967	4459308	17.16

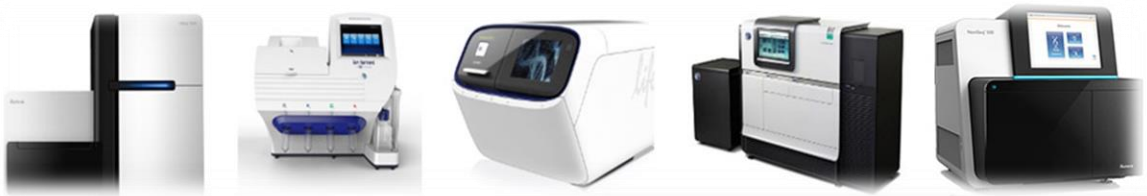
Joshua Quick, University of Birminham, Feb 15, 2018

- Lower throughput
- Higher cost per GB than NGS
- Lower sequencing quality then NGS
- Hard calling genetic variants and mutations

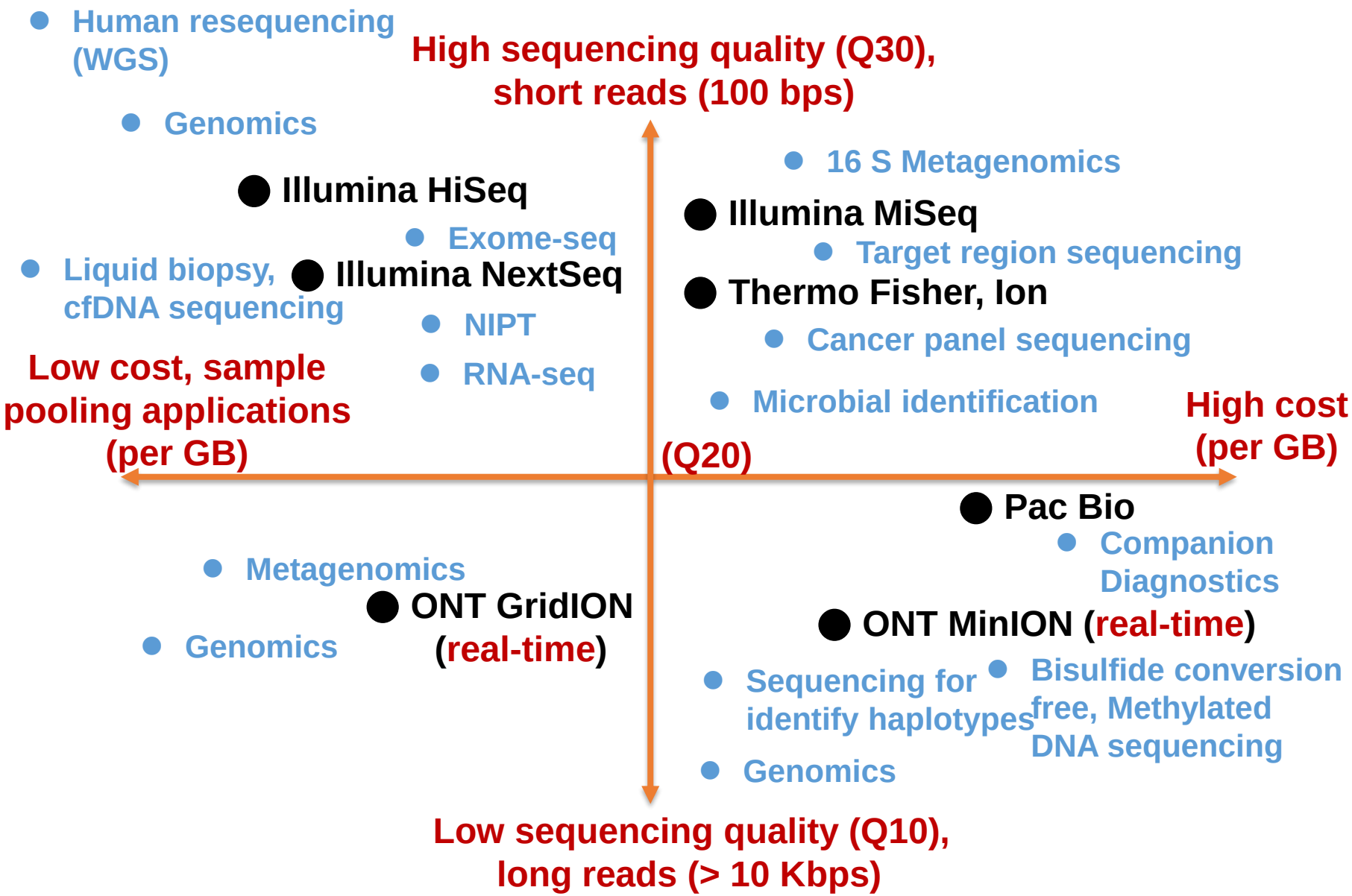


Clive G. Brown, 2016

- For medical applications
- For academic research applications
- Key questions
  - How to choose platforms?
  - Will it be appropriate to integrated two platforms for an application?



Strategy for Selecting Seq. Platforms



Application-1

Exploring Human Genome Structural Variants (SV)

Human SV – Current Knowledge

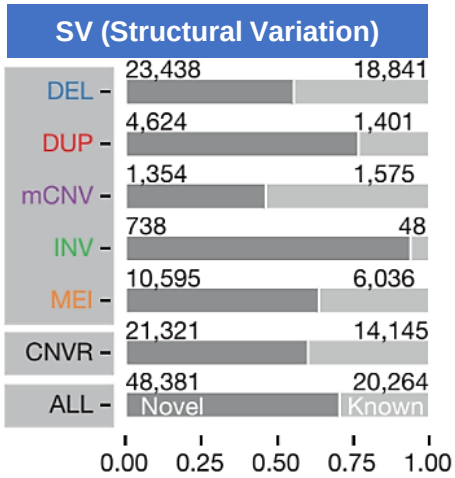
ARTICLE

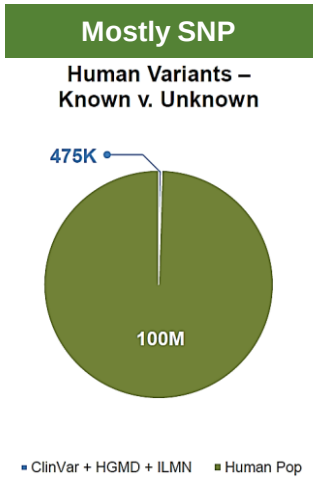
OPEN

doi:10.1038/nature15394

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.





~ 4 K
~ 18.4 Mbp

Individual Level:

~ 4 M
~ 4 Mbp

Nature. 2015 Oct 1;526(7571):75-81.

Catalog of Human SV:

- Illumina WGS (~100 bp reads, mean 7.4X coverage)
- 26 human populations

"…although SNPs contribute more eQTLs overall, our results suggest that SVs have a disproportionate impact on gene expression relative to their number."

"We further expand the number of candidate SVs in strong LD with GWAS hits by ~30% (39/136 novel associations implicating SVs as candidates) and find that GWAS haplotypes are enriched up to threefold for common SVs, which emphasizes the relevance of ascer-taining SVs in disease studies."

"…while many SVs in our callset are statistically phased, the diploid nature of the genome is non-optimally captured by current analysis approaches, which mostly rely on mapping to a haploid reference. We envision that in the future, the use of technology allowing substantial increases in read lengths over the current state-of-the-art will enable genomic analyses of truly diploid sequences to facilitate targeting these additional layers of genomic complexity."

Human Genome Assembly by ONT Reads

ARTICLES

nature biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30x theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. *De novo* assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). We developed a protocol to generate ultra-long reads (N50 > 100 kb, read lengths up to 882 kb). Incorporating an additional 5× coverage of these ultra-long reads more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4-Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38.

Nat Biotechnol. 2018 Apr;36(4):338-345.

The Final Assembled Genome:

- 41 FC (Regular), 5 FC (Ultra-long)
- 91.2 Gb, ~30X
- Ultra-long Reads, N50 > 100 Kb, up to 882 Kb

- 2,867 million bases
- Covering 85.8% of reference genome
- Assembly accuracy: 99.8% (incorporating complementary short-read data)

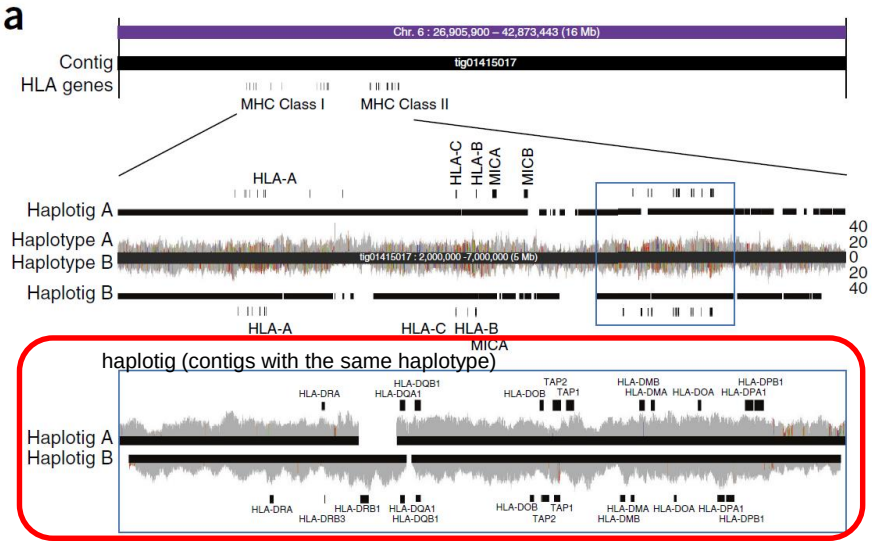
- NG50 ~3 Mb
- With Ultra-long reads, NG50 ~6.4 Mb
- With Ultra-long reads, achieved phasing of the entire 4 Mb MHC locus

Ultra-long Reads, Assembly and Telomeres

nature biotechnology

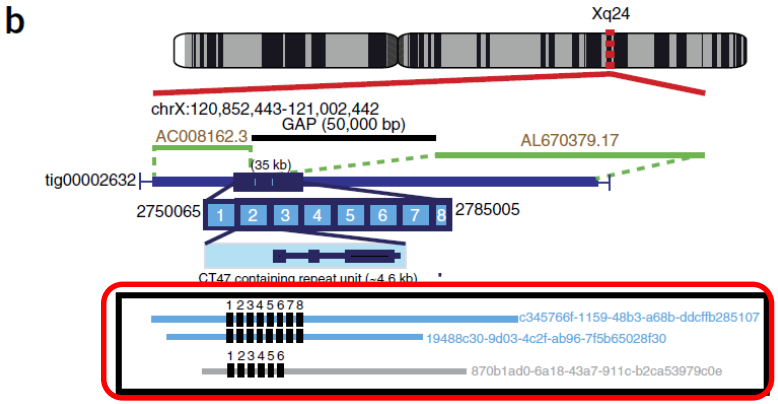
OPEN

- Heterozygous SNPs were called using Illumina data
- Phased using the ultra-long nanopore reads
- Generate two **pseudo-haplotypes**

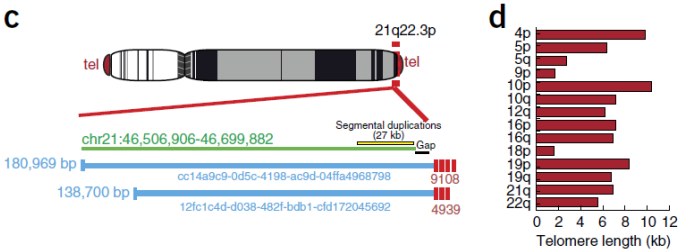


- A 16-Mbp ultra-long read contig and associated haplotigs are shown spanning the full MHC region.
- The first time the MHC has been assembled and phased over its full length in a diploid human genome.

Nat Biotechnol. 2018 Apr;36(4):338-345.



Two reads provide evidence for an array of eight repeat copies and one read supports six copies, suggesting heterozygosity.

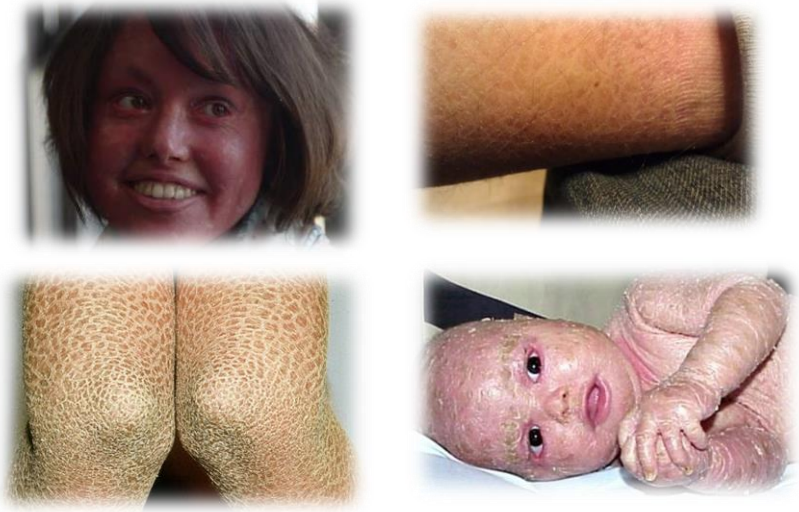


- FISH (fluorescent in situ hybridization) estimates and direct cloning of telomeric DNA suggests that telomere repeats (**TTAGGG**) extend for multiple kbs at the ends of each chromosome.
- Evidence for telomeric arrays that span 2–11 kb within 14 subtelomeric regions for GM12878.

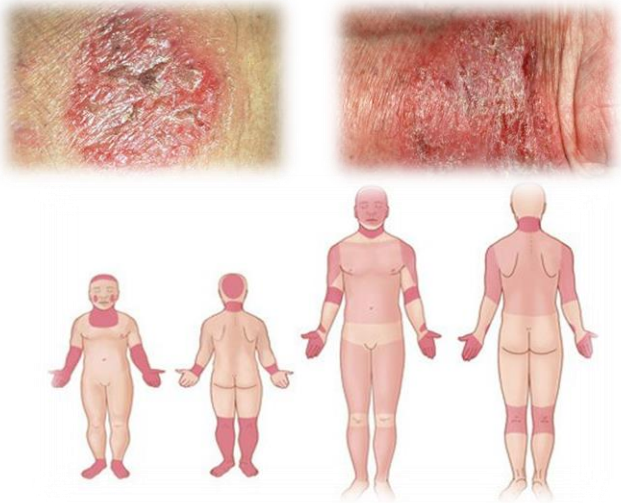
One remarkable molecule: Filaggrin

- Ichthyosis vulgaris (尋常魚鱗病); Eczema (濕疹)
- The CNV allele frequencies in the Irish population were found to be 33.9% **10** repeats, 51.5% **11** repeats and 14.6% **12** repeats. Shortest genotype (10,10): Eczema risk: 1.67
- When null mutations are excluded, each additional filaggrin repeat gained decreases the odds ratio for atopic eczema by 0.88.
- Filaggrin CNV makes a significant, **dose-dependent** contribution to eczema risk

Ichthyosis vulgaris (魚鱗病)

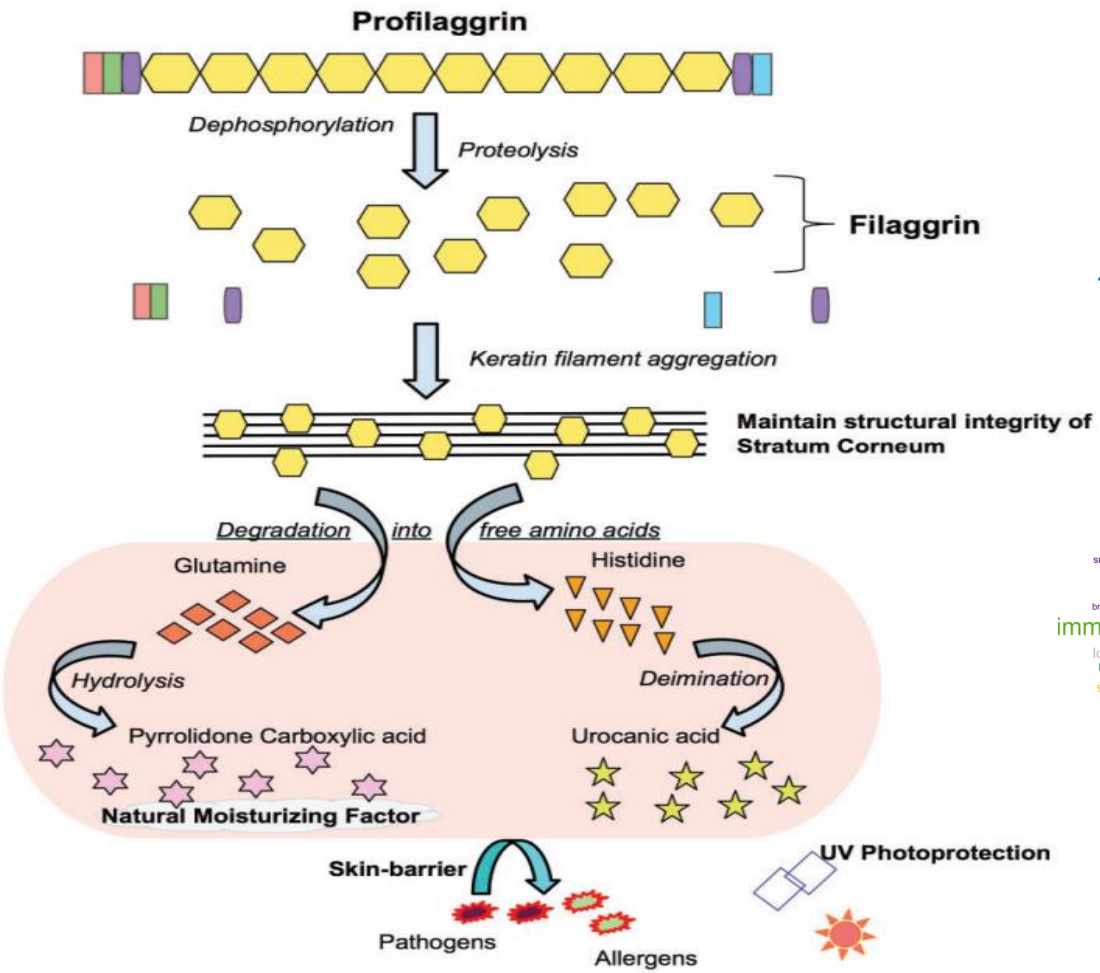


Eczema (濕疹)



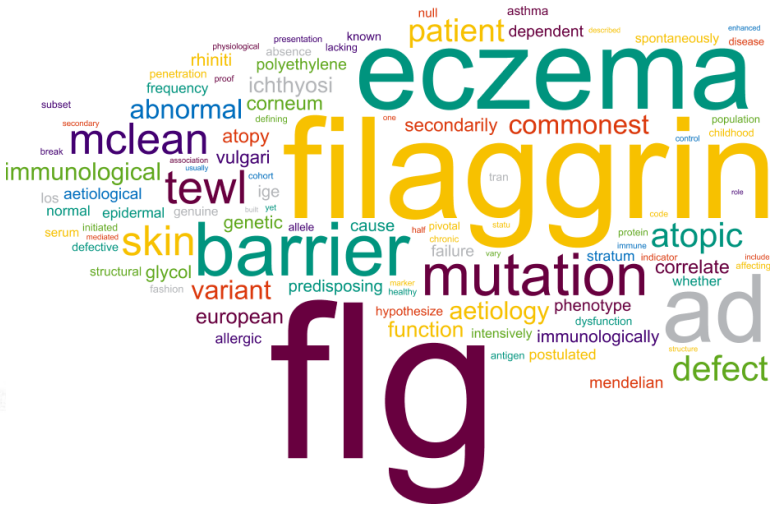
J Invest Dermatol. 2012

Processing and function of Profilaggrin



FLG helps upkeep the structural integrity of the skin in mammals.

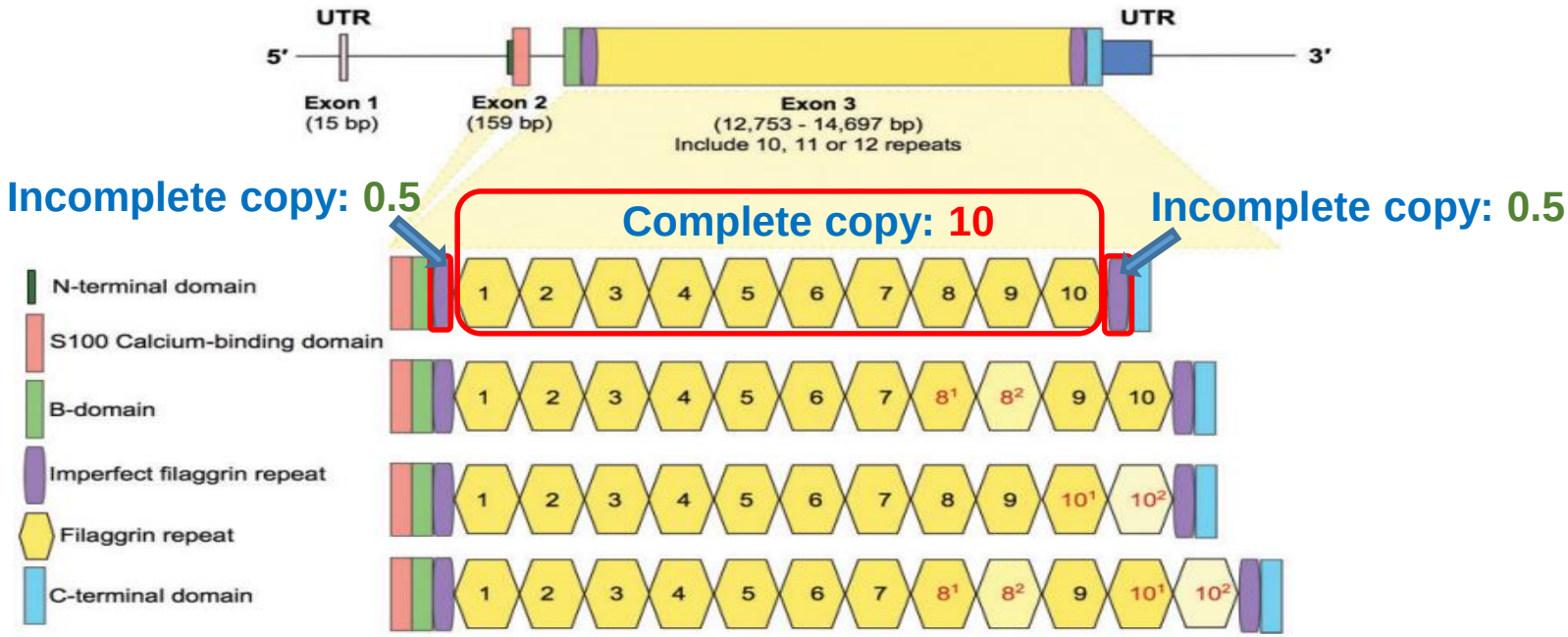
Filaggrin were shown to guard skin from UV and contribute to the natural moisturizing factor of the skin.



http://www.researchperspectives.org

Genome Biol. Evol. 2016, 8(10):3240–3255

Structure of FLG and Copy number



- Filaggrin repeats were shown to be copy number variable ranging from 10 to 12 copies among human populations.
- The copy number of these repeats was negatively associated with atopic dermatitis susceptibility.
- However, a comprehensive documentation of the global distribution of FLG genetic variation free of ascertainment bias has yet to be compiled.

Genome Biol. Evol. 2016, 8(10):3240–3255

Tandem Repeats Variations

COMMENTARY

See related article on pg 98

Profilaggrin, Dry Skin, and Atopic Dermatitis Risk: Size Matters

John A. I

Mutation
risk facto
gies. Nev
in *FLG* al
findings i
utility in i

Journal of In

been suggested that the number of FLG repeats can relate to a dry skin phenotype; i.e., fewer repeats may lead to less FLG protein expression and drier skin (Ginger *et al.*, 2005). What is currently not known, however, is whether there is a

J Invest Dermatol. 2012 Jan;132(1):10-1.

Pediatric Dermatology Vol. 34 No. 3 e140–e141, 2017

Intragenic Copy Number Variation in the Filaggrin Gene in Ethiopian Patients with Atopic Dermatitis

Abstract: Gene
involving truncati
number variation a
of developing atop
Asian populations
have been identifi
ation between *FL*
severity in a small
proposed. We stu
copy number and
association betwe
ber, suggesting th
factors are of m
Ethiopians to AD.

long-range polymerase chain reaction (PCR) amplification of the *FLG* repeats, which has previously been described for various populations, including Africans (4). The total *FLG* CN was classified as low (20–21) or high (22–24).

A total of 105 cases were successfully genotyped and phenotyped and included in the study. The long-range PCR yielded products of sizes that confirmed a different

Pediatr Dermatol. 2017 May;34(3):e140-e141.

Research letter

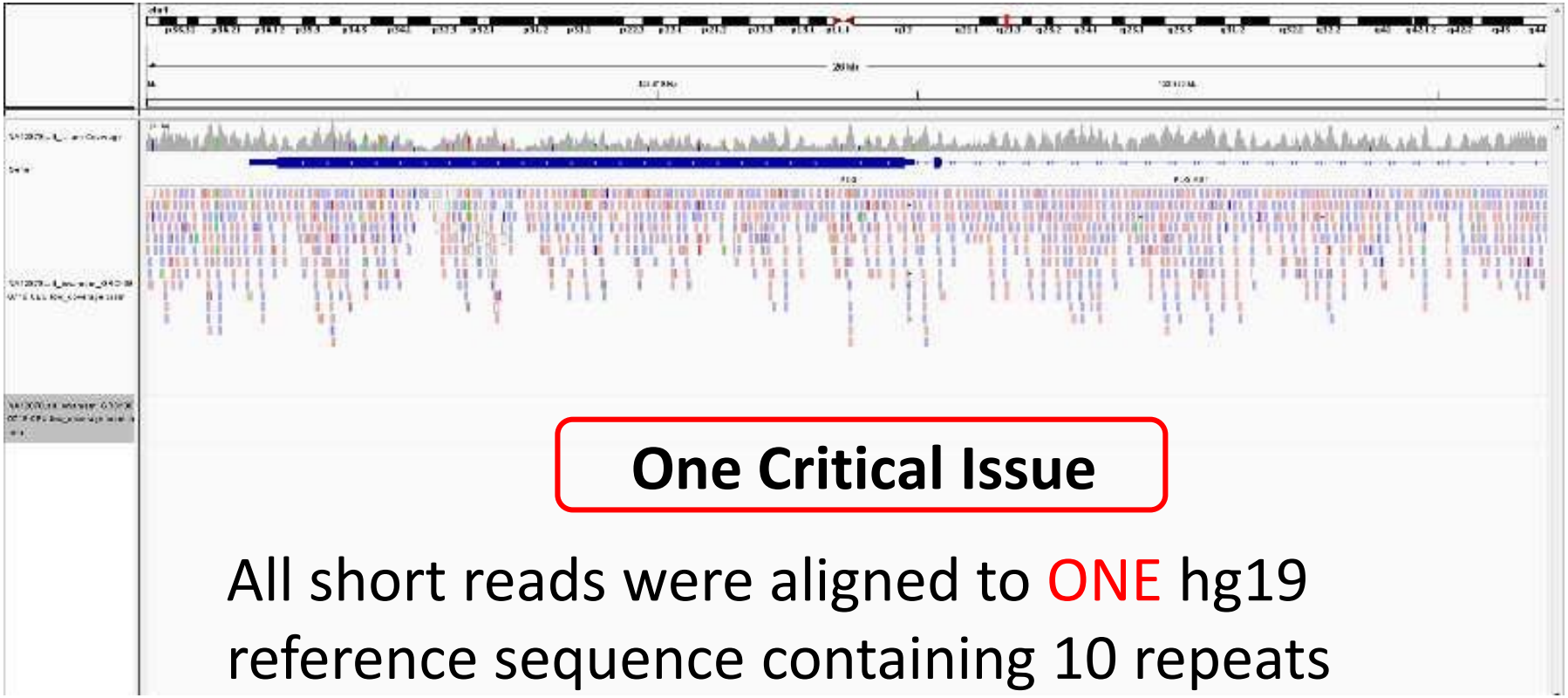
Copy-number variation of the filaggrin gene in Korean patients with atopic dermatitis: what really matters, 'number' or 'variation'?

DOI: 10.1111/bjd.14287

Dear Editor, Since the articles reporting a methodological breakthrough on the full sequencing of the gene encoding profilaggrin (*FLG*),[1,2] associations between loss-of-function mutations of *FLG* and atopic dermatitis (AD) have been reported across ethnicities.[3,4] However, both the low prevalence of *FLG* mutations in patients with AD in some nations (< 4% in Italy) and the high prevalence of *FLG* mutations in healthy control in other nations (~ 10% in Ireland) suggest that factors other than *FLG* mutation may be at work.[5,6] Brown *et al.* introduced an interesting new factor contributing to the risk of AD: copy-number variation aa(CNV).[6] *FLG* is polymorphic, with allelic variants of 10–12 nearly identical repeats in exon 3.[6] They

Br J Dermatol. 2016 May;174(5):1098-100.

Conventional Short Reads Alignment



One Critical Issue

All short reads were aligned to ONE hg19 reference sequence containing 10 repeats

Cannot make high confidence variant calls in the repeat region using short reads

Long Range PCR for full length of Filaggrin

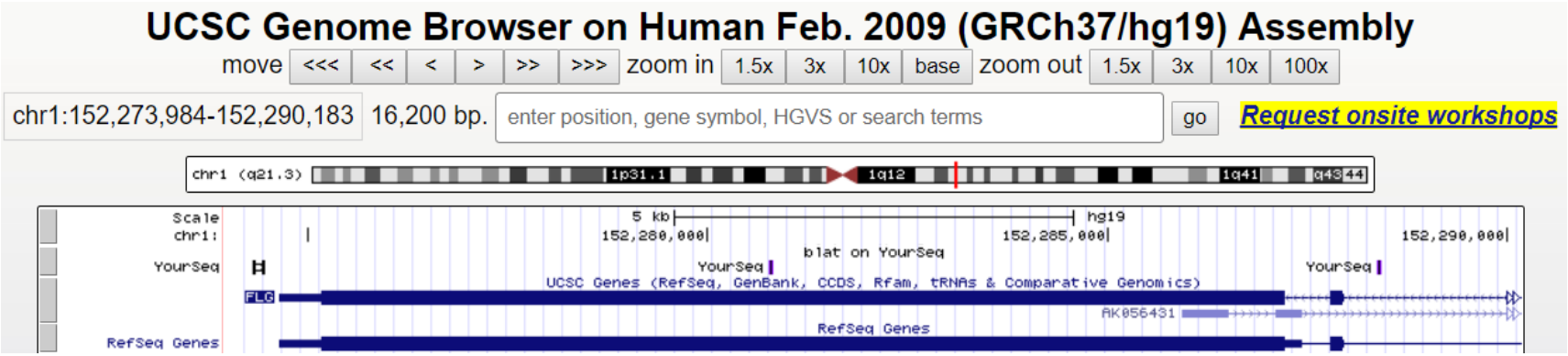
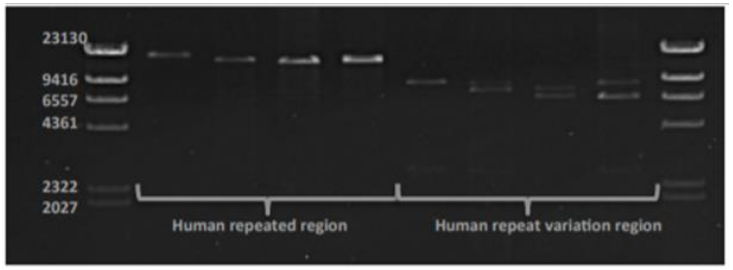
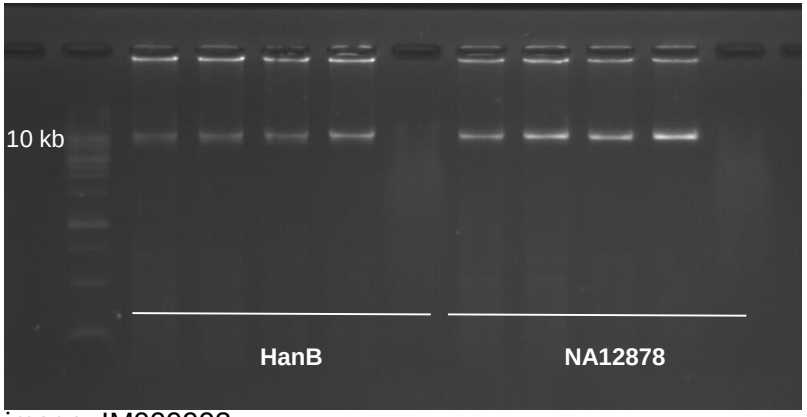


Table S1. Primer sets for the filaggrin gene repeat region in human, chimpanzee, gorilla, orangutan, crab-eating macaque, and human variation in repeated region.

REPEATED REGION	
Human	**hFLG (13969 bp in total)**
Forward	5′ -CTTGTCATATGGCTAACTGGCTTTCAGAGA-3′
Reverse	5′ -ATTGTGGGACAGTGATTATGTTGGAGAAAA-3′
Human variation in repeated region	**hFLGv (6480 bp in total)**
Forward	5′ -GTGCAAGCAGAAAAACATATGACA -3′
Reverse	5′ -CCTGTTTCGTGATCTGCCTTTGACATGG -3′



PCR Amplification



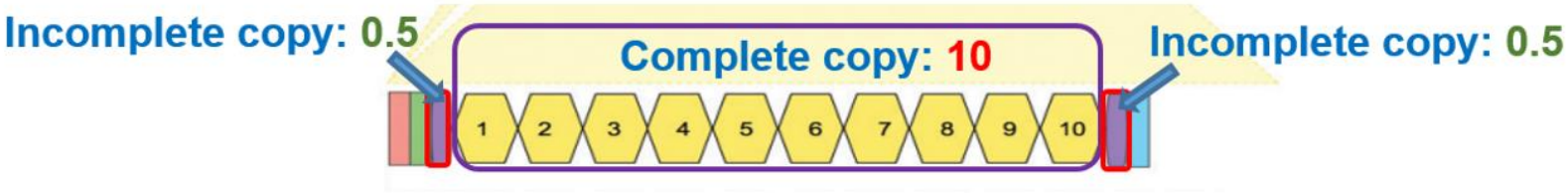
image_IM009993

Romero V et al BMC Evolutionary Biology 17:10, 2017

Preliminary Results

Sample	HanB1	HanB2	HanB3	NA12878
Reads (Q>=7)	3,182	6,246	49,346	277850
Align to FLG ref	4262	5068	34558	9169
>10,000 bp	686	724	3762	2443
Both F and R found	323	445	1780	1533
TRN Determined	(11+1/12+1)	(11+1/12+1) (with noise)	(11+1/12+1)	(10+1/12+1)

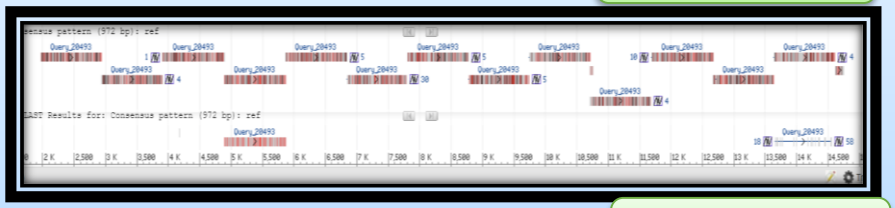
10 + 0.5 + 0.5 = 10 + 1



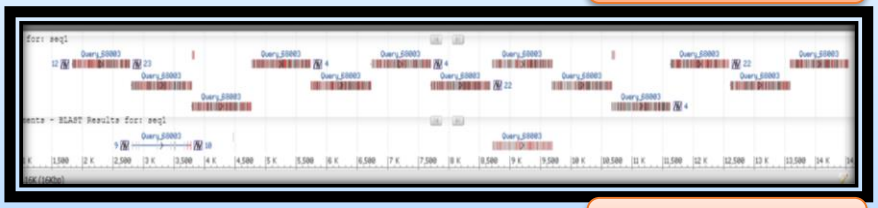
Identifying the copy number of FLG gene

Blast result

12+1 copy

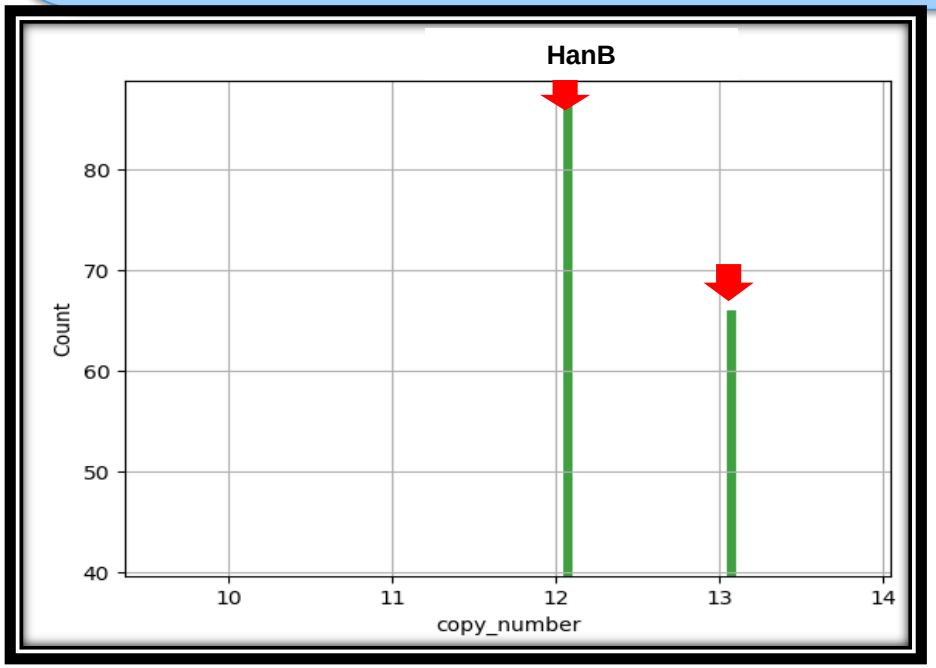


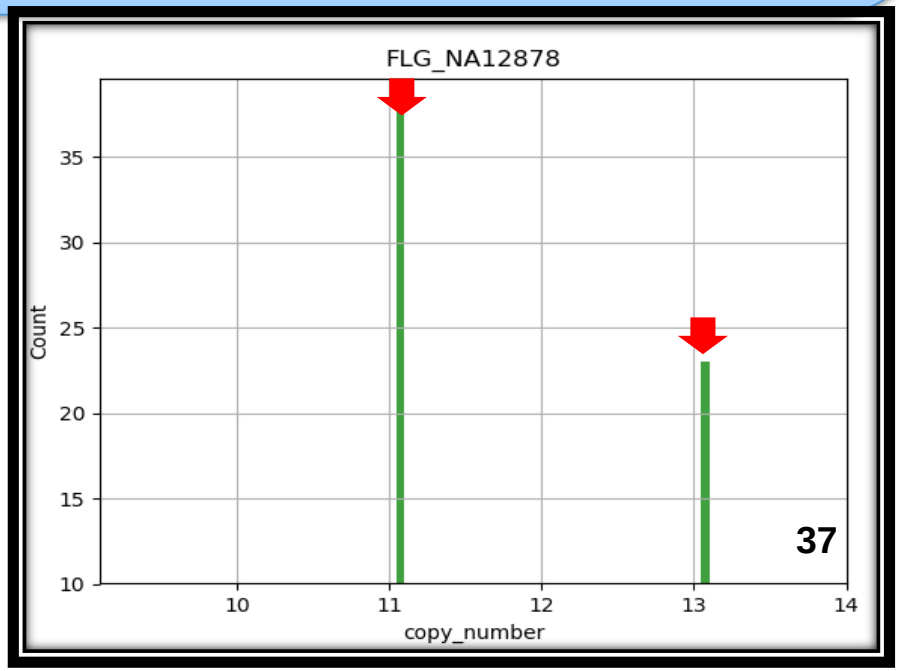
12+1 copy



11+1 copy

10+1 copy



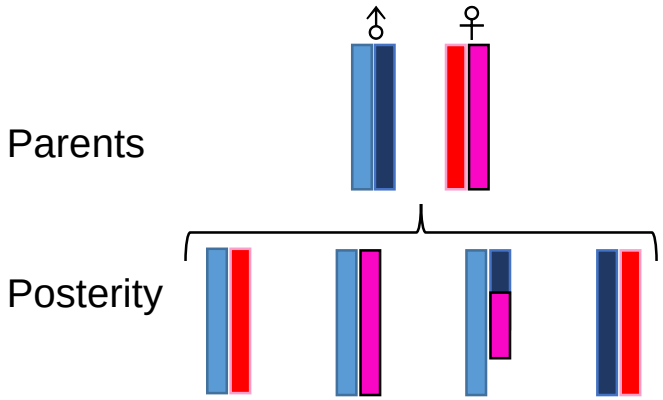


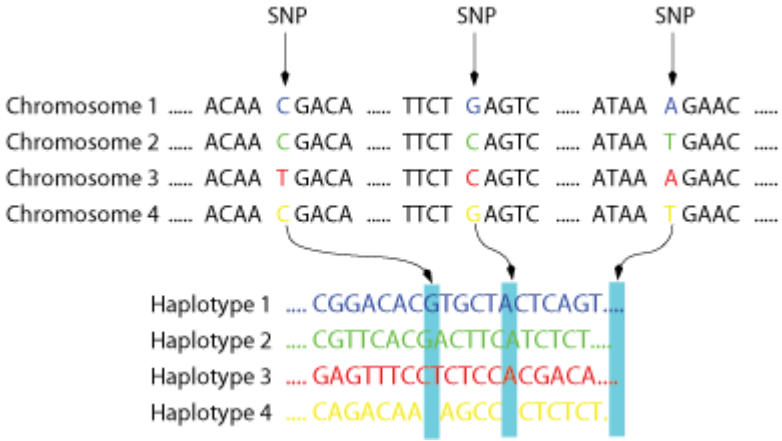
Application-2

Full genotyping for high polymorphic loci in Pharmacogenomics

Haplotypes (Full genotyping of a high polymorphic locus)

- A haplotype is a group of genes within an organism that was inherited together from a single parent
- The term "**haplotype**" can also refer to the inheritance of a cluster of single nucleotide polymorphisms (SNPs)
- There are highly polymorphic (more haplotypes) on genes of CYP or HLA

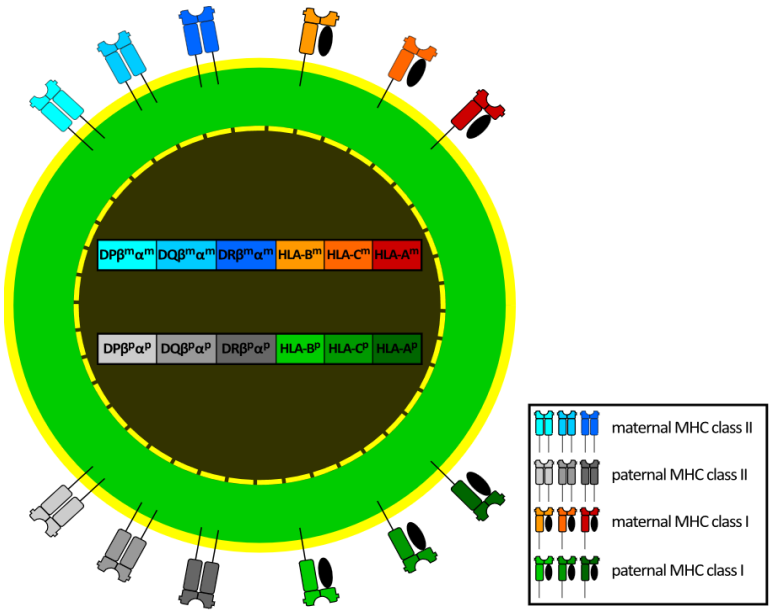




Filippo Geraci and Marco Pellegrini

Human Leukocyte Antigen (HLA)

- The **h**uman **l**eukocyte **a**ntigen (HLA) genes are the human versions of the **m**ajor **h**istocompatibility **c**omplex (MHC)
- HLA system is the locus of genes that encode for proteins on the surface of cells that are responsible for regulation of the immune system in humans

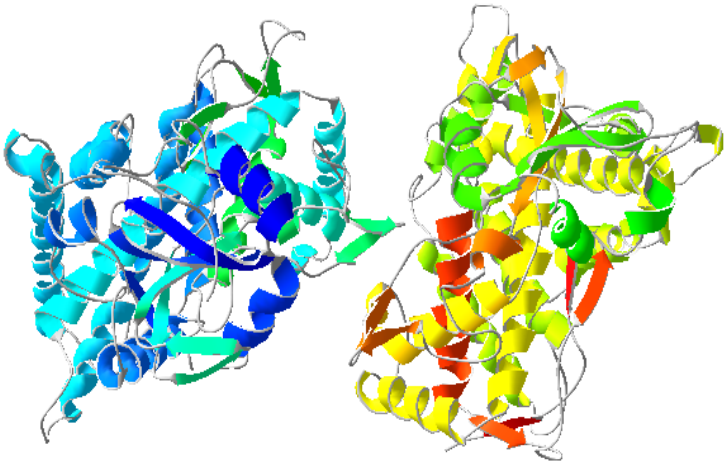


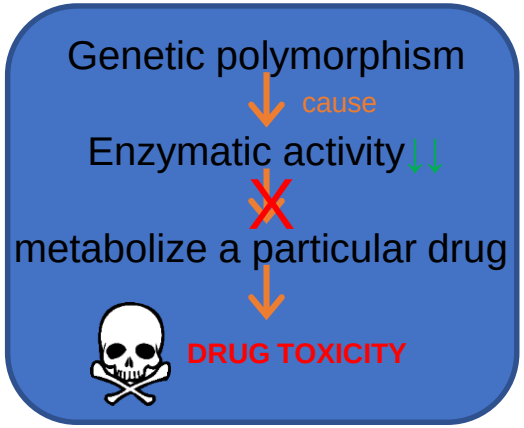
Gene	Number of Alleles (Haplotypes)
HLA-A	4,081
HLA-B	4,950
HLA-C	3,685
HLA-DQA1	94
HLA-DQB1	1,178
HLA-DPA1	64
HLA-DPB1	963
HLA-DRB1	2,146

Nucleic Acids Research. 2015 43:D423-431

Cytochrome P450 Enzymes

- Cytochrome P-450 mixed-function oxidase (CYP)
  - abundant in the liver and other organs
  - responsible for the metabolism of many drugs and environmental chemicals (antiarrhythmics, adrenoceptor antagonists, tricyclic antidepressants)
- Genetic polymorphisms of CYP2D6, CYP2C9, and CYP2C19 have been well studied



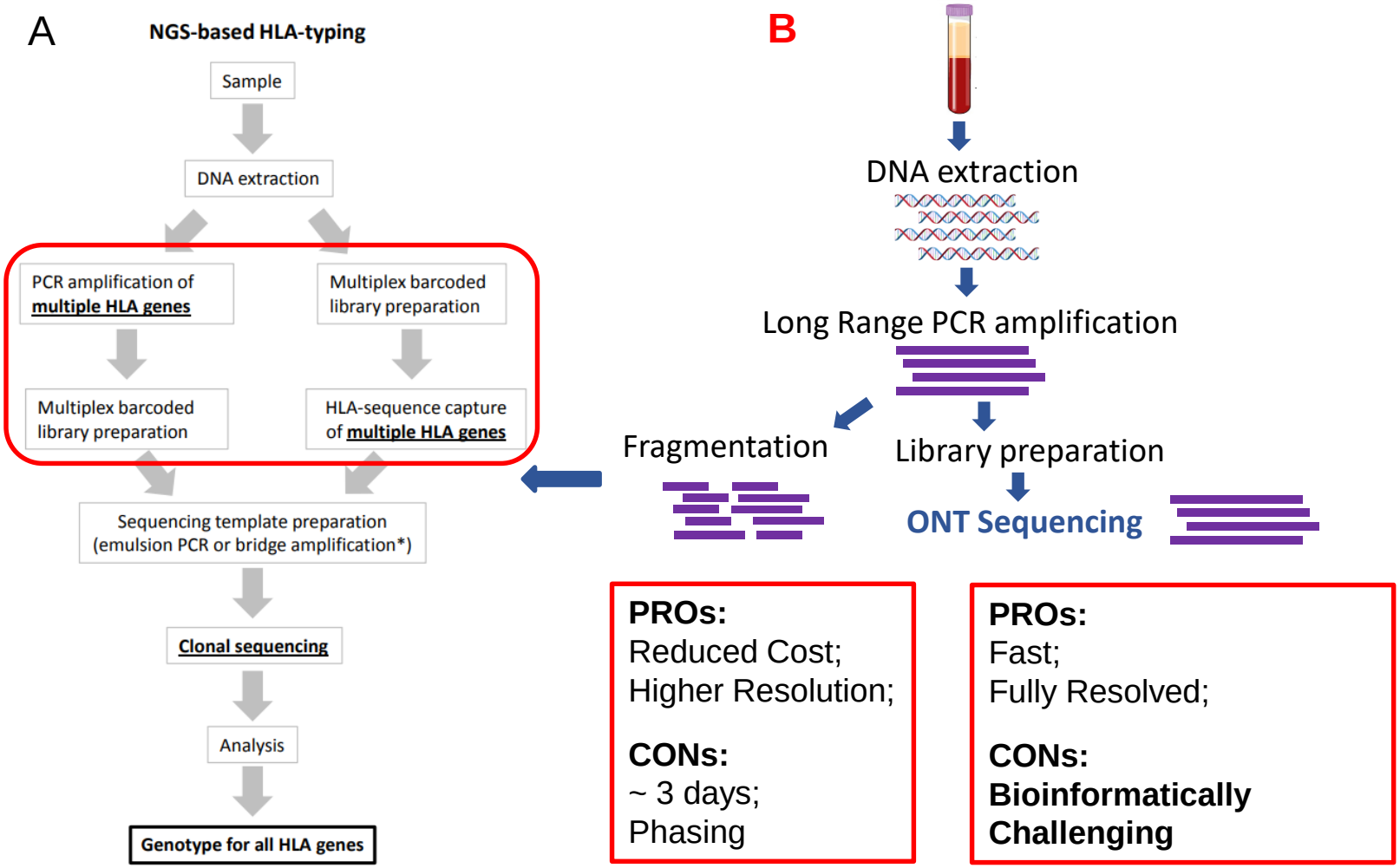


Examples of Haplotype-Drug Associations

Gene	Haplotypes or genotypes	Phenotypes	Related Drug	Clinical Interpretation
ATM	WT/WT	rs11212617 CC genotype	Metformin	NORMAL RESPONSE
CYP1A2	*1A/*1F	Ultrarapid Metabolizer	Cyclobenzaprine	INCREASE DOSE
CYP2A6	*1/*1	Normal Metabolizer	Nicotine	NORMAL RESPONSE
CYP2C19	*1/*2	Intermediate Metabolizer	Sertraline	USE CAUTION
CYP2C9	*1/*1	Normal Metabolizer	Warfarin	NORMAL DOSE
CYP2D6	*4/*35	Intermediate Metabolizer	Codeine	CONSIDER ALTERNATIVES
CYP3A4	*1A/*1B	Intermediate Metabolizer	Alfentanil	DECREASE DOSE
CYP3A5	*1A/*3A	Expresser	Sirolimus	NORMAL RESPONSE
CYP4F2	*1/*1	Normal Metabolizer	Phenprocoumon	NORMAL RESPONSE
DDRGK1	WT/c.510+364T>G	rs6051639 AC genotype	Ribavirin	USE CAUTION
DPYD	*5/*9A/c.496A>G/IVS10-15T>C	Normal Metabolizer	Capecitabine	NORMAL RESPONSE
F2	WT/WT	Wild Type	Oral-Contraceptives	NORMAL RESPONSE
F5	WT/WT	Non Factor V Leiden Carrier	Eltrombopag	NORMAL RESPONSE
G6PD	WT/WT	Normal G6PD Efficiency	Chlorpropamide	NORMAL RESPONSE
HLA-B	15:02/08:20	15:02	carbamazepine	CONSIDER ALTERNATIVES

https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm

Platform for full genotyping of HLA loci



Hum Immunol. 2016 Nov;77(11):1016-1023.

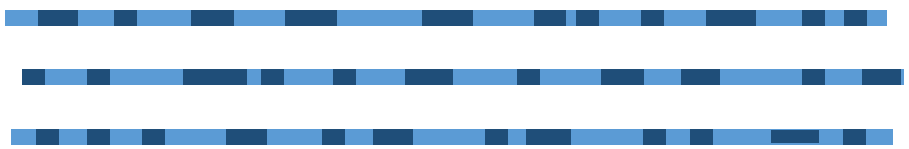
High polymorphic genes associated with drugs efficacy and adverse events

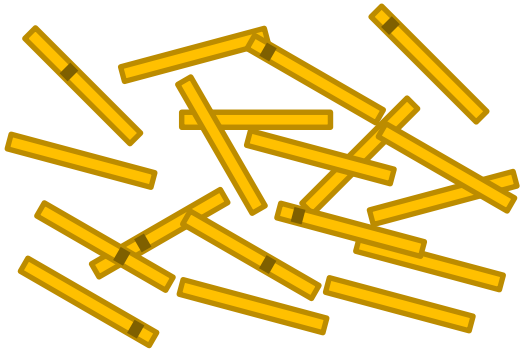
U.S. FOOD & DRUG ADMINISTRATION

Pharmacogenomic Biomarkers in Drug Labeling					
ALK	CYP2B6	F2	HPRT1	NPM1	ROS1
BCHE	CYP2C19	F5	IFNL3	PDGFRA	SERPINC1
BCR-ABL1	CYP2C9	FIP1L1-PDGFRA	IL12A	PDGFRB	SLCO1B1
BRAF	**CYP2D6**	FLT3	IL12B	PGR	TPMT
BRCA	CYP3A5	G6PD	IL23A	PML-RARA	TPP1
CASR	DMD	GALNS	IL2RA	POLG	**UGT1A1**
CD274	DPYD	**HLA-A**	KIT	PROC	VKORC1
CFTR	EGFR	**HLA-B**	MS4A1	PROS1	
CYB5R	ERBB2	**HLA-DQA1**	MYCN	RAS	
CYP1A2	ESR	**HLA-DRB1**	NAGS	RET	

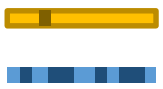
https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm

Long read & Short read



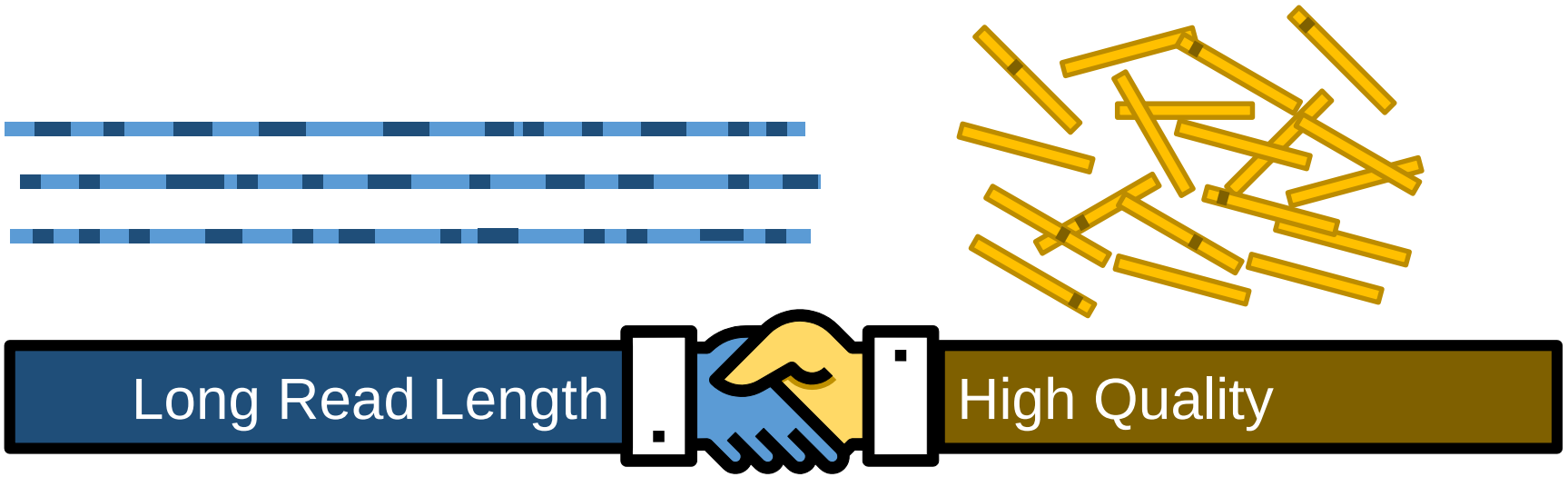


ONT		NGS
low (Q10)	Quality	Very High (Q30)
Very Long	Read Length	short
1kb ~ 882kb[1]		0.5 kb

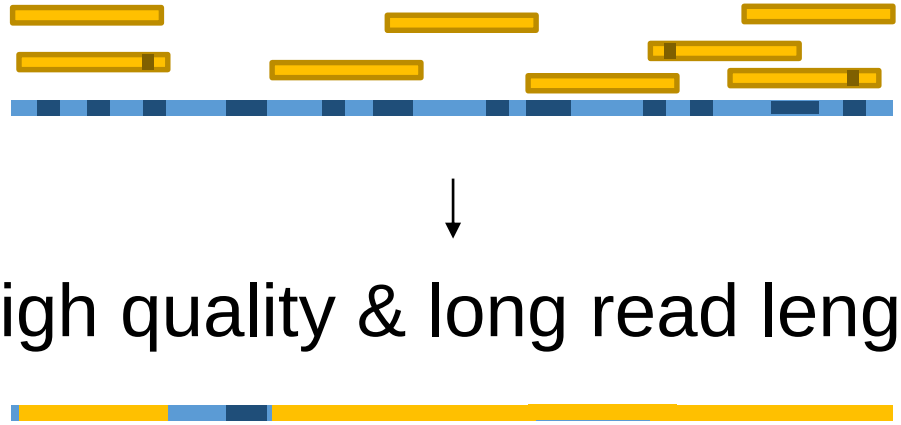


The deep color region meant bad quality region.

[1]Nature Biotechnology 36, pages 338–345 (2018).

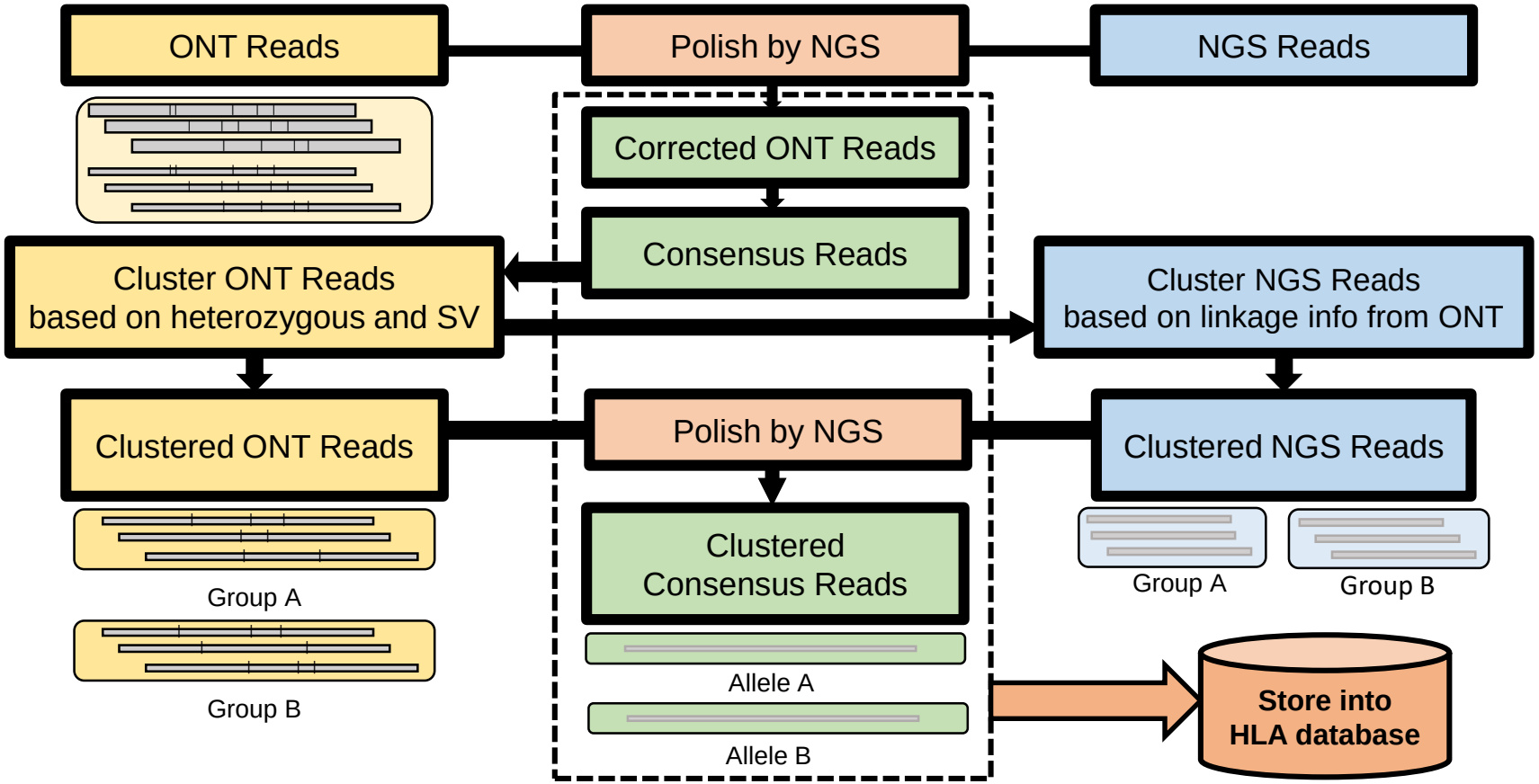


Hybrid correction (Polished)

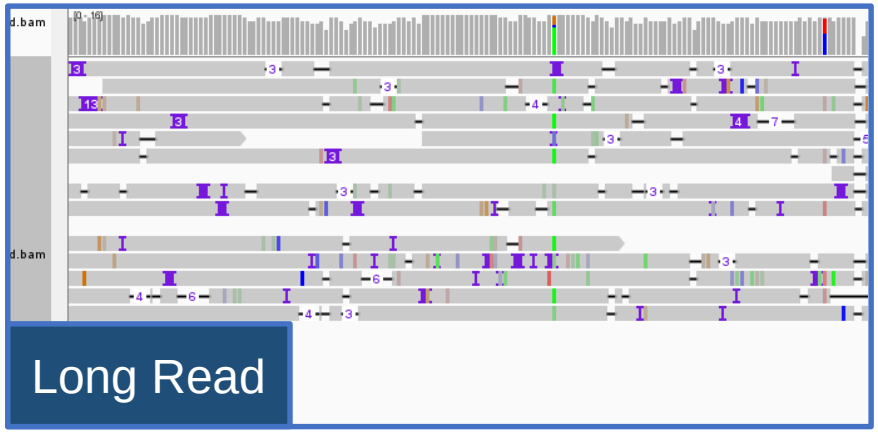


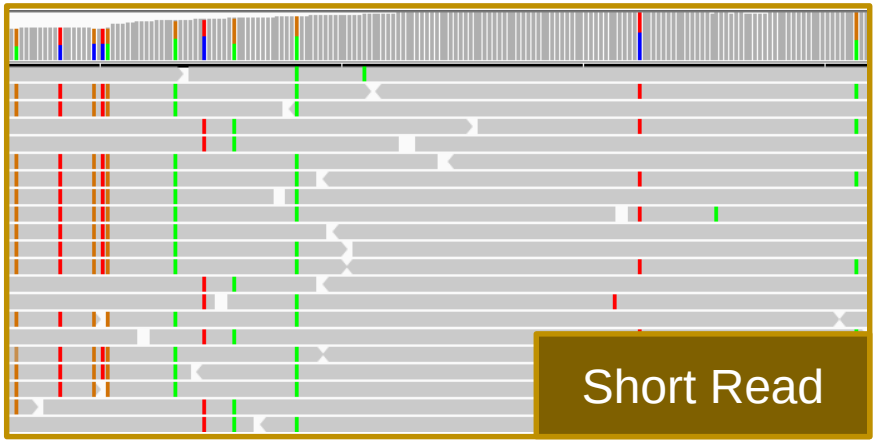
High quality & long read length

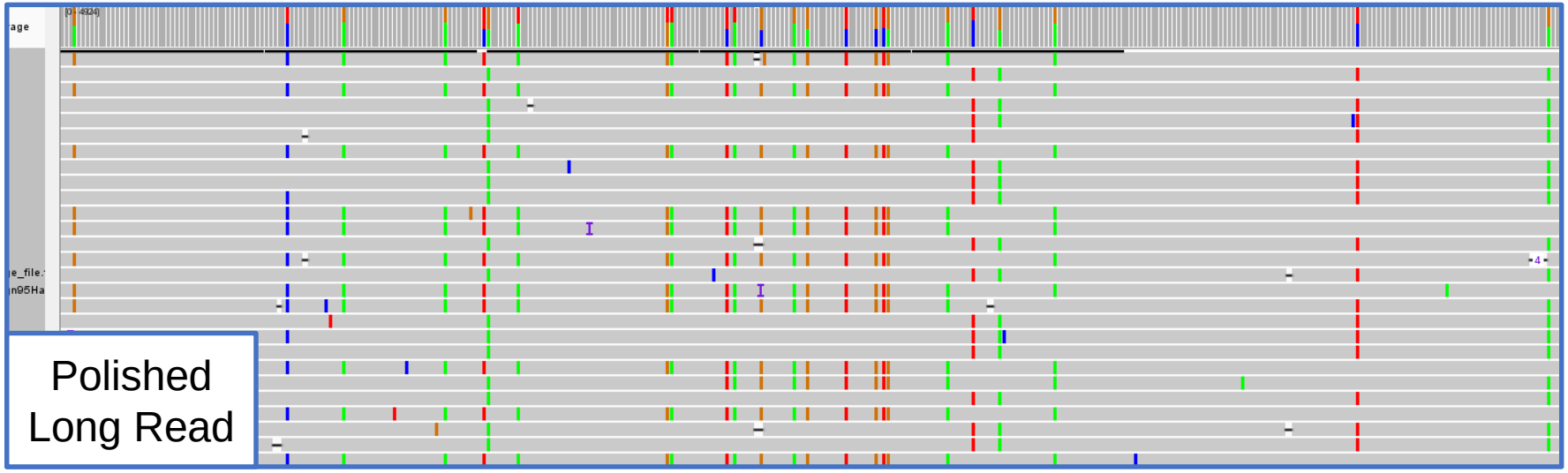
Sequencing Analysis Work Flow

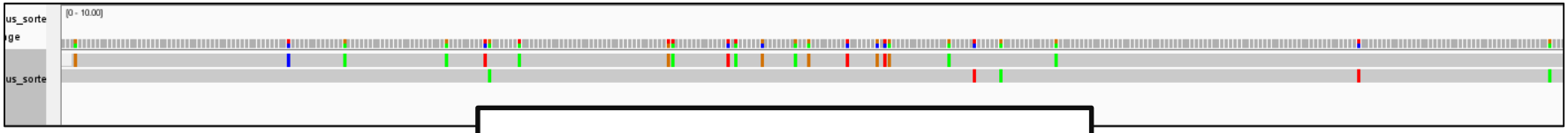


Construction of HLA-A alleles of HanB









Consensus allele sequences

Top alignment to IMGT: The allele type of HLA-A, HLA-B and HLA-C form HanB

HLA-A*02:10

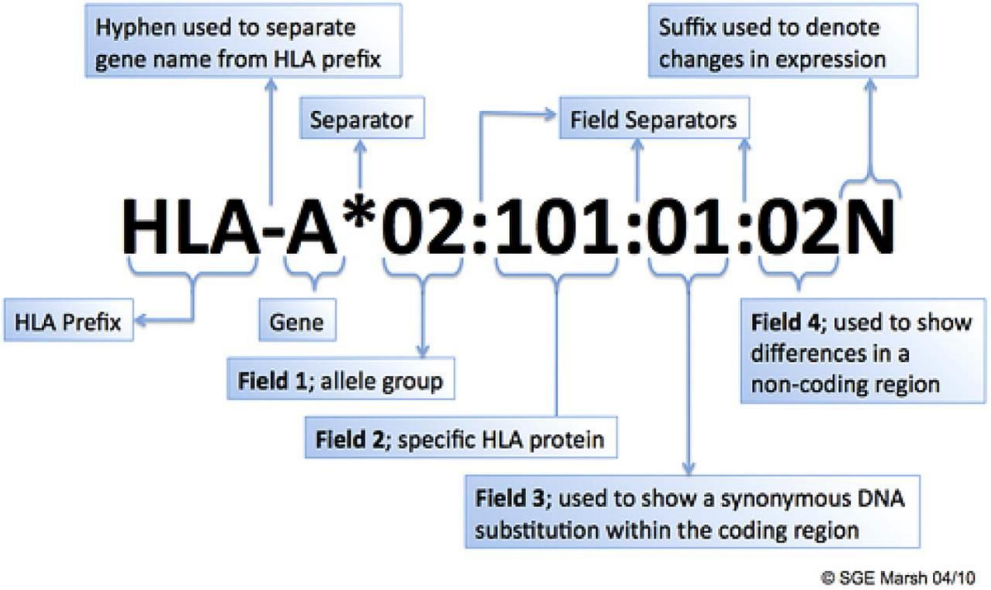
HLA-A*24:02:01:01

HLA-B*15:32:01

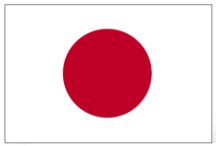
HLA-B*40:06:01:01

HLA-C*04:01:01:14

HLA-C*08:01:01:01



Constructing Database in Different Populations



SCIENTIFIC REPORTS

OPEN

Mapping the genetic diversity of HLA haplotypes in the Japanese populations

Received: 17 July 2015
Accepted: 06 November 2015
Published: 09 December 2015

Woei-Yuh Saw[1,2], Xuanyao Liu[1,3], Cl
Tomohiro Katsuya[6], Ryosuke Kimu
Ken Yamamoto[11], Mitsuhiro Yokota
Yik-Ying Teo[1,2,3,4,5,13,*] & Norihiro Ka

*"Several studies[47–49] have reported the risk of inaccuracies and confounding in genetic association studies in populations even with relatively small genetic differences. In this line, based on our data, we can further advocate caution in using a generic Japanese panel (e.g., JPT in the HapMap) for **imputation of SNPs** and HLA alleles in samples from Okinawa Prefecture."*

Sci Rep. 2015 Dec 9;5:17855.

27:1597–1607 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

Assembly and analysis of 100 full MHC haplotypes from the Danish population

Jacob M. Jensen,[1] Palle Villesen,[1,2] Rune M. Friborg,[1] The Danish Pan-Genome Consortium,[5] Thomas Mailund,[1] Søren Besenbacher,[1,3] and Mikkel H. Schierup[1,4]

[1]Bioinformatics Research Centre, Aarhus University, 8000 Aarhus C., Denmark; [2]Department of Clinical Medicine, Aarhus University, 8200 Aarhus N., Denmark; [3]Department of Molecular Medicine, Aarhus University Hospital, Skejby, 8200 Aarhus N., Denmark; [4]Department of Bioscience, Aarhus University, 8000 Aarhus C., Denmark

*"we reconstruct full MHC haplotypes from de novo assembled trios…. We report 100 full MHC haplotypes and call a large set of structural variants in the regions for future use in **imputation with GWAS data**."*

Genome Res. 2017 Sep;27(9):1597-1607.

Journal of Clinical & Cellular Immunology

Gowda et al., J Clin Cell Immunol 2016, 7:2
http://dx.doi.org/10.4172/2155-9899.1000399

Research Article Open Access

Comparative Analyses of Low, Medium and High-Resolution HLA Typing Technologies for Human Populations

Malali Gowda[1,2*], Sheetal Ambardar[1], Nutan Dighe[3], Ashwini Manjunath[1], Chandana Shankaralingu[1], Pradeep Hirannaiah[1], John Harting[4], Swati Ranade[4], Latha Jagannathan[5*] and Sudhir Krishna[5*]

[1]Next Generation Genomics Laboratory, Centre for Cellular and Molecular Platform, National Centre for Biological Sciences, TIFR Bangalore, India

*"This is the first case study of HLA typing using second and third generation NGS technologies for an Indian population. The PacBio platform is a promising platform for large-scale HLA typing for **establishing an HLA database for the untapped ethnic populations of India**."*

J Clin Cell Immunol 2016, 7:2

Significant Implications of HLA Haplotyping

nature COMMUNICATIONS

MHC matching improves engraftment of iPSC-derived neurons in non-human primates

Asuka Morizane[1], Tetsuhiro Kikuchi[1], Takuya Hayashi[2], Hiroshi Mizuma[2], Sayuki Takara[2], Hisashi Doi[2], Aya Mawatari[2], Matthew F. Glasser[3], Takashi Shiina[4], Hirohito Ishigaki[5], Yasushi Itoh[5], Keisuke Okita[6], Emi Yamasaki[1], Daisuke Doi[1], Hirotaka Onoe[2,7], Kazumasa Ogasawara[5], Shinya Yamanaka[6,8] & Jun Takahashi[1,9]

Nat Commun. 2017 Aug 30;8(1):385.

Common Allele/Haplotyping

BIOINFORMATION
Discovery at the interface of physical and biological sciences

Bioinformation. 2017; 13(3): 94–100. PMCID: PMC5450251
Published online 2017 Mar 31. doi: 10.6026/97320630013094

T-cell epitopes predicted from the Nucleocapsid protein of Sin Nombre virus restricted to 30 HLA alleles common to the North American population

Sathish Sankar,[1,*] Mageshbabu Ramamurthy,[1] Balaji Nandagopal,[1] and Gopalan Sridharan[1]

Bioinformation. 2017; 13(3): 94–100.

Allelic Expression

G3 Genes | Genomes | Genetics

Unique Allelic eQTL Clusters in Human MHC Haplotypes

Tze Hau Lam,* Meixin Shen,* Matthew Zirui Tay,† and Ee Chee Ren*,‡,1
*Singapore Immunology Network, A*STAR, Singapore 138648, †Department of Molecular G
Duke University, Durham, North Carolina 27710, and ‡Department of Microbiology and Imm
School of Medicine, National University of Singapore, Singapore 117597

Immunotherapy(Antibody based and Cell based)

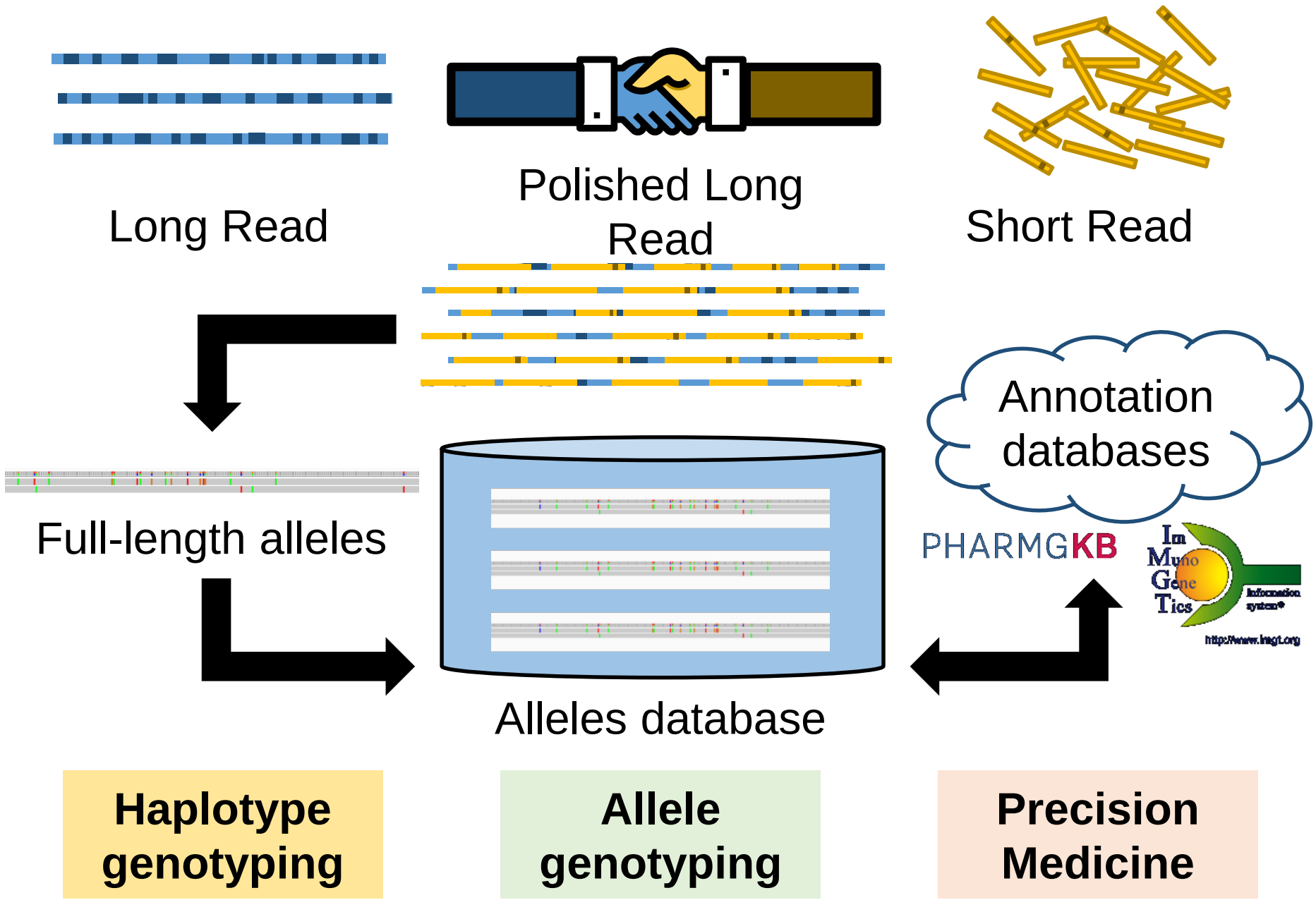
Science REPORTS

Cite as: D. Chowell *et al.*, *Science* 10.1126/science.aao4572 (2017).

Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy

Diego Chowell,[1,2] Luc G.T. Morris,[2,3*] Claud M. Grigg,[4*] Jeffrey K. Weber,[5] Robert M. Samstein,[1,2] Vladimir Makarov,[1,2] Fengshen Kuo,[1,2] Sviatoslav M. Kendall,[1,2] David Requena,[6] Nadeem Riaz,[1,2,7] Benjamin Greenbaum,[8] James Carroll,[9] Edward Garon,[9] David M. Hyman,[10,14] Ahmet Zehir,[11] David Solit,[1,10,12] Michael Berger,[1,11,12] Ruhong Zhou,[5,13] Naiyer A. Rizvi,[4†] Timothy A. Chan[1,2,7,14†]
[1]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [2]Immunogenomics and Precision Oncology Platform,

Science. 2018 Feb 2;359(6375):582-587.


Long Read
Polished Long Read
Short Read
Full-length alleles
Annotation databases
PHARMGKB
Alleles database
Haplotype genotyping
Allele genotyping
Precision Medicine

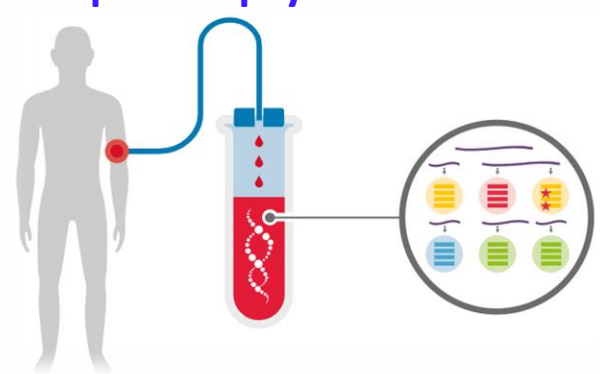
Sequencing Platform Comparison

Platform Comparison	Illumina	Thermo Fisher	Pacific Biosciences	Oxford Nanopore
Sequencing by Synthesis	Yes	Yes	Yes	No
DNA Size Selection	Yes	Yes	Yes	No
Post-library Amplification	Yes	Yes	No, Single Molecule	No, Single Molecule
Detection	Fluorescent Imaging	Ion Semiconductor	Fluorescent Imaging	Ionic Current Change
Sequencing Rate (s/base)	2 – 20 sec	30 sec	0.25 sec	0.002 sec
Running Time	Fixed	Fixed	Fixed	Run & Stop
DNA Sequencing	Yes	Yes	Yes	Yes
Direct DNA Modification Detection	No	No	Yes	Yes
Direct RNA Sequencing	No	No	No	Yes
Read Length	Short, up to 300 bp X 2	Short, up to 600 bp	Long, Average 6-8 Kb	Long, Average 6-30 Kb
Total Reads (M)	4 – 800 (PE)	2 – 130	0.3 – 0.5	0.3 – 0.5
Total Base (Gb)	1.2 - 120	0.3 – 25 /Chip	5 – 8 /SMRT Cell	2 – 10 /Flow Cell
Instrument Cost (USD)	20K – 275K	200 – 300K	350K	1K – 125K

Applications Requiring Throughput (coverage)

- Liquid biopsy for diagnosis and prognosis
- Liquid biopsy for reproductive and genetic health
- Human genome resequencing
- Signal counting applications, such as gene expression profiling
- Detection of DNA mutation (diagnostic testing)

Liquid biopsy



Resequencing

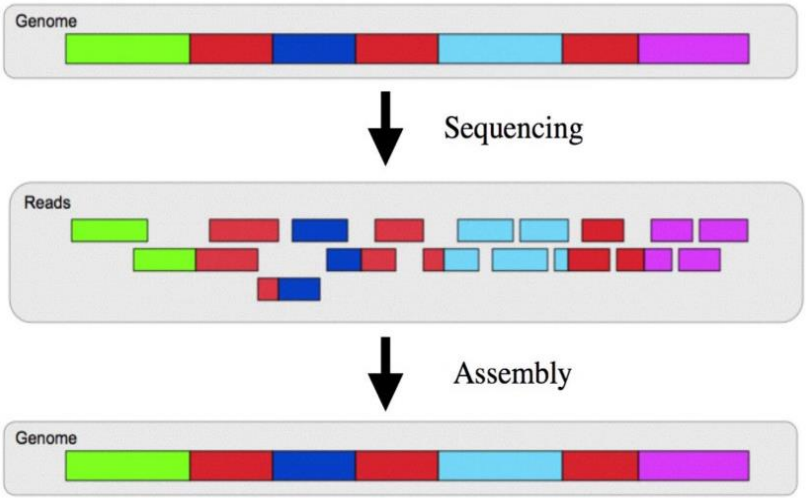
Coverage

https://www.inivata.com/liquid-biopsy/

Applications Requiring Long Reads

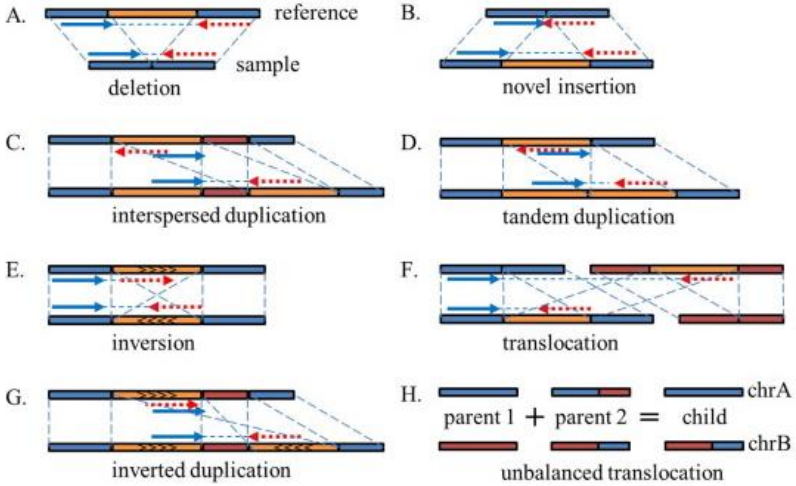
- Genome assembly (New resolution)
- Detection of structural variations
- Full genotyping of highly polymorphic loci

Genome assembly



https://www.youtube.com/watch?v=5wvGapmA5zM

Structural variation detection



Methods. 2016 Jun 1;102:36-49.

Thank you for your attention

Long

Short