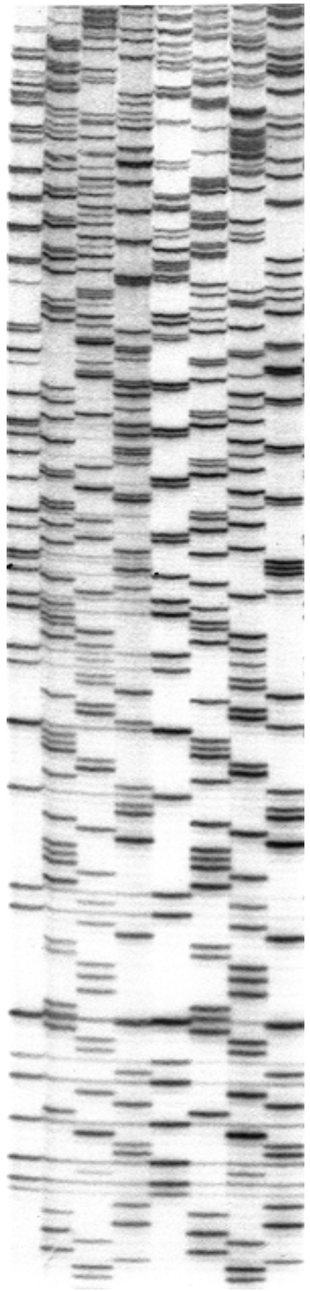
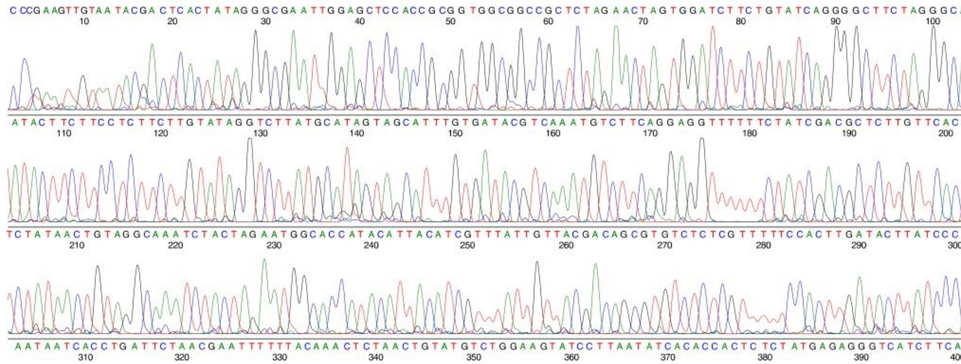
Analiza genomu

Od sekwencji do diagnozy

Sekwencjonowanie tradycyjne (Sangera) - pierwszej generacji

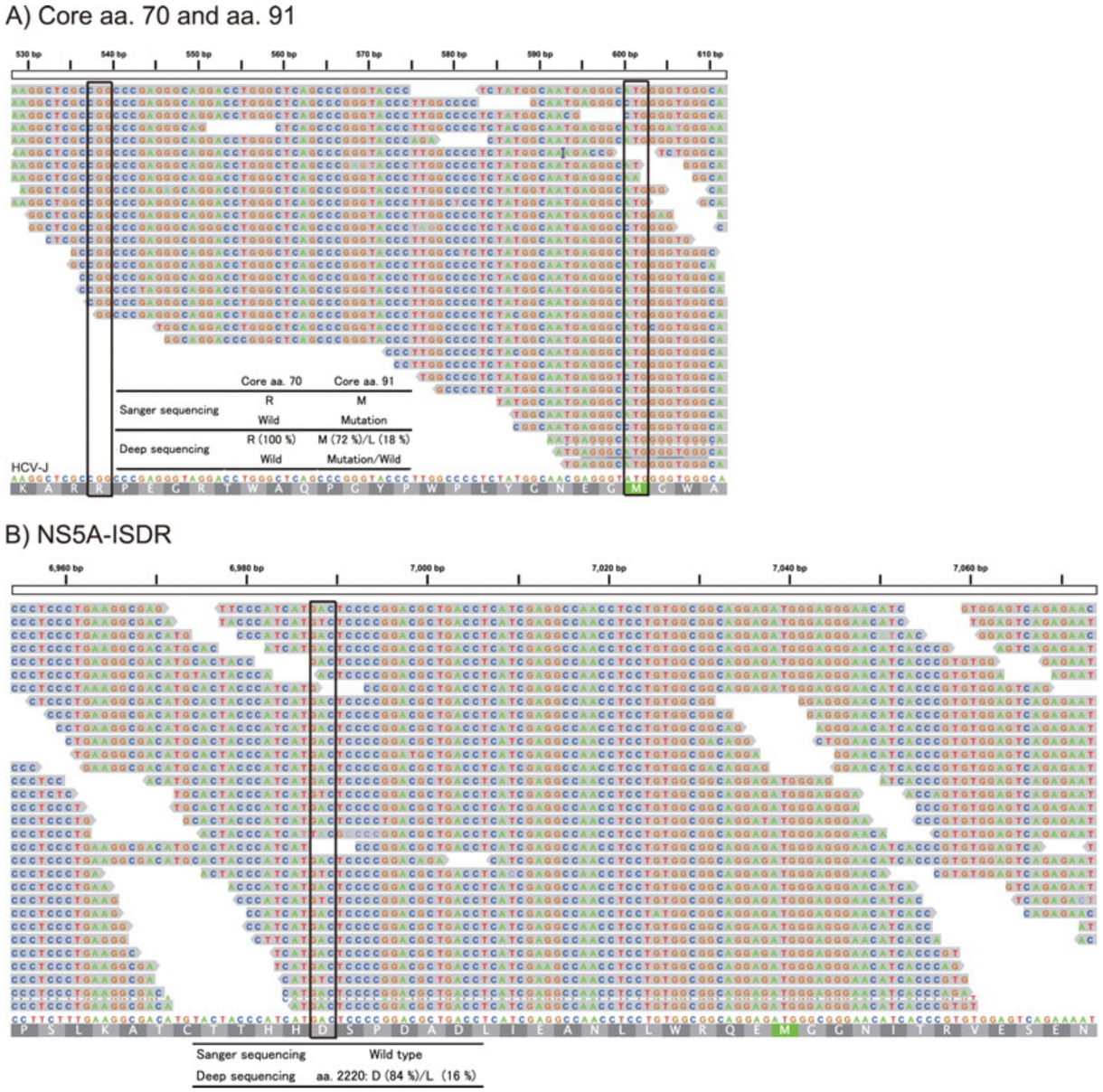
- Sanger, 1977 (konkurencyjna metoda Maxama i Gilberta się nie przyjęła)
- Pierwsze automatyczne sekwenatory kapilarne (ABI Prism) w 1986
- Sekwencjonowanie przez syntezę
- Odczyty do 800 zasad, typowo ~600
- Niska przepustowość (96 odczytów/ urządzenie/przebieg)
- Wciąż standard dla krótkich sekwencji i małych zadań





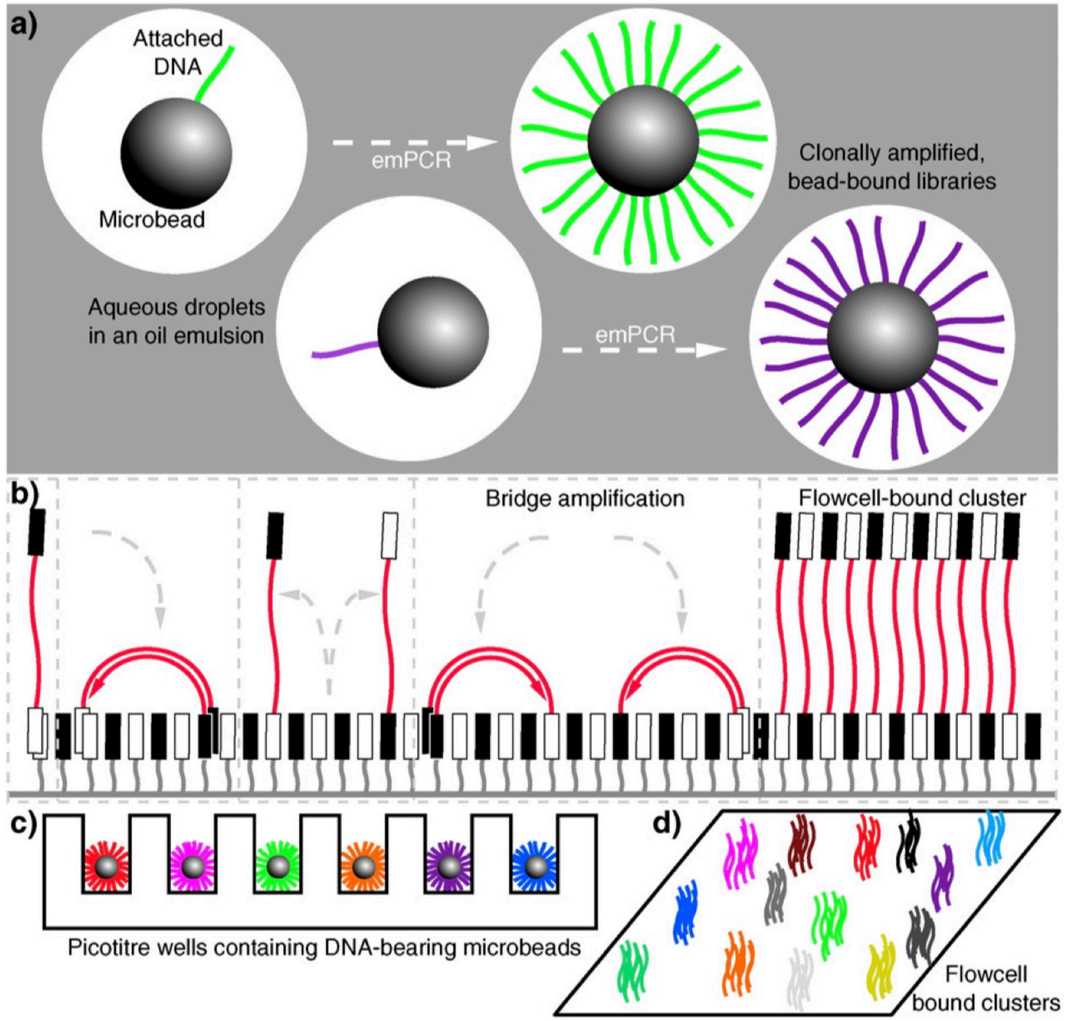
Sekwencjonowanie nowej (drugiej) generacji - NGS

- Od 2005 (454)
- Tzw. głębokie sekwencjonowanie (deep sequencing) albo sekwencjonowanie masywnie równoległe
- Sekwencjonowanie przez syntezę
- Miliony odczytów w jednym przebiegu do 10^{12} zasad
- Odczyty krótkie: 50-400 zasad



NGS drugiej generacji

- Obecnie dwie główne platformy
- Illumina
  - Amplifikacja "mostkowa" na podłożu stałym
  - Sekwencjonowanie przez syntezę - odwracalne terminatory i detekcja fluorescencyjna
- Ion Torrent (Life Technologies)
  - Biblioteki namnażane przez PCR w emulsji
  - Sekwencjonowanie przez syntezę, detekcja zmian pH po uwalnianiu protonów przy dodaniu nukleotydu



The sequence of sequencers: The history of sequencing DNA

James M. Heather *, Benjamin Chain

Genomics 107 (2016) 1–8

Sekwencjonowanie

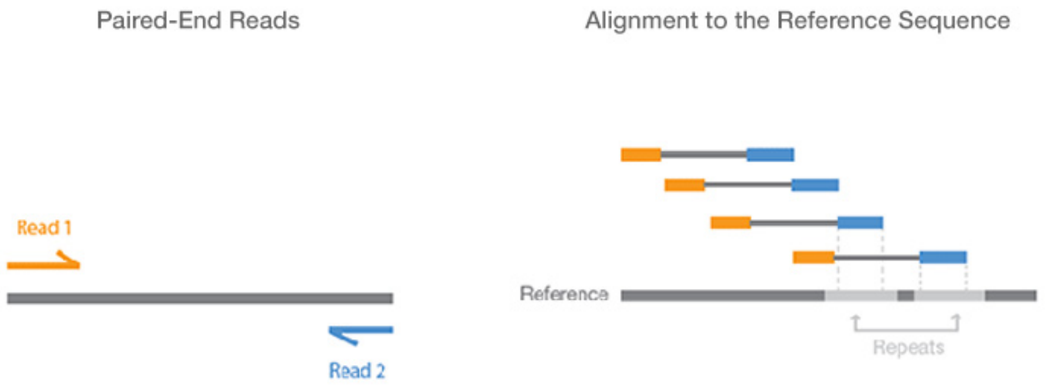
- Głównym wyzwaniem w sekwencjonowaniu nie jest sam odczyt sekwencji
- Odczytywane fragmenty są krótkie
  - do ~700-800 nt (sekw. tradycyjne Sangera)
  - 50-400 nt (NGS)
- Wyzwaniem jest złożenie długiej sekwencji z tych krótkich fragmentów
  - różne technologie sekwencjonowania wymagają różnych metod obróbki bioinformatycznej

Wyzwania NGS

- Narzędzia bioinformatyczne ciągle rozwijane
- Mapowanie odczytów do znanej sekwencji referencyjnej ułatwia zadanie, ale nadal są problemy
  - sekwencje repetytywne
  - zmienność strukturalna
  - faza haplotypu
- NGS to nie jest "czarna skrzynka"

Sparowane odczyty (paired ends)

- Sekwencje z obu końców dłuższego (kilka kb) fragmentu
- Ułatwia zmapowanie



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

© illumina.com

NGS ukierunkowane (targeted)

- Wzbogacenie o wybrane obszary genomu
  - amplifikacja (PCR)
  - wychwytywanie przez hybrydyzację
- WES - Whole Exome Sequencing - tylko obszary kodujące
- panele wybranych genów

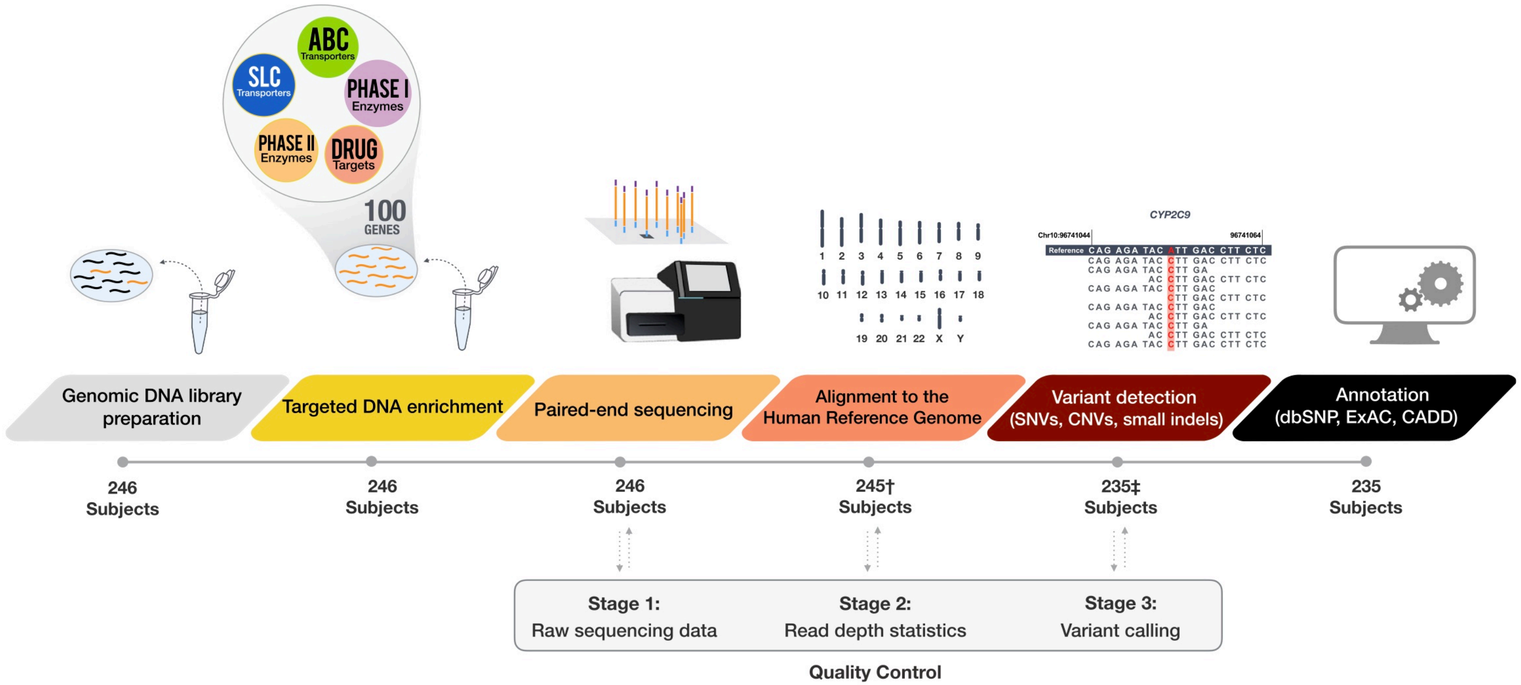
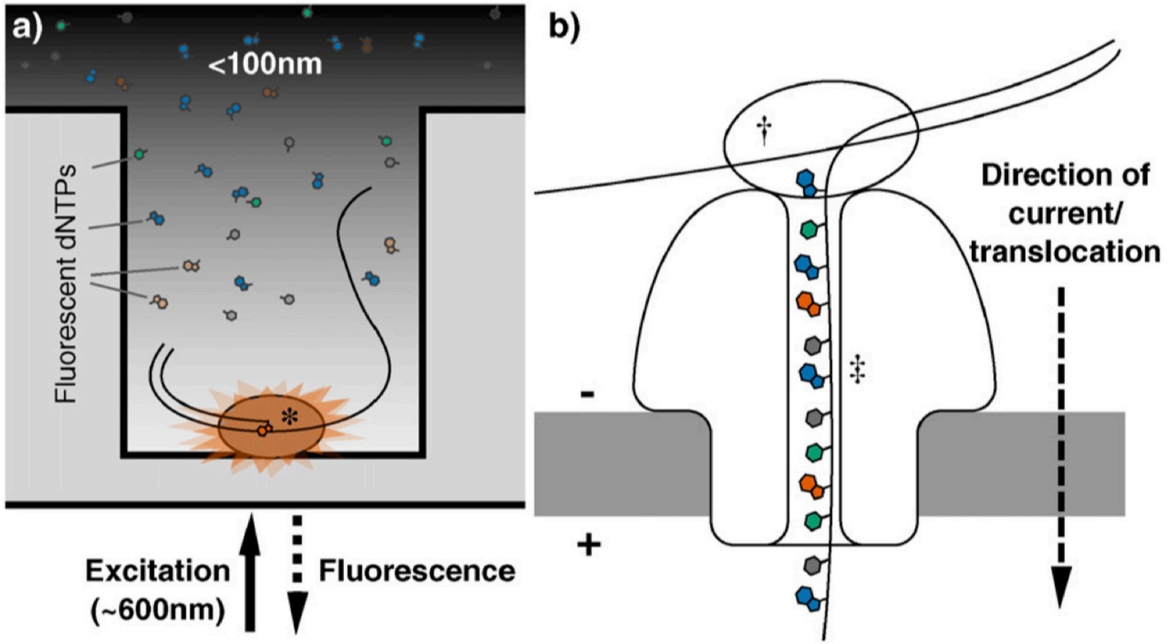


Fig. 1 PGxSeq sample and data processing workflow ($n = 246$). Eleven subjects were excluded from variant analysis due to low read count (†; $n = 1$) and high GC content (‡; $n = 10$). All clip art depicted in this Figure has been created by the authors

Gulilat *et al. BMC Medical Genomics* (2019) 12:81
https://doi.org/10.1186/s12920-019-0527-2

NGS trzeciej generacji

- Sekwencjonowanie z długimi odczytami
- Obecnie dwie główne platformy
  - single-molecule real time sequencing (SMRT) - Pacific Biosystems
  - nanopore sequencing (MinION) - Oxford Nanopore
- Odczyty 20-30 kb



The sequence of sequencers: The history of sequencing DNA

James M. Heather *, Benjamin Chain

Genomics 107 (2016) 1–8

Sekwencjonowanie III generacji

- Metody (PacBio, Nanopore) dające długie odczyty (>10 000 nt)

Powerful

Get up to 50 Gb data from a single flow cell*.

* Theoretical max output when system is run for 72 hours at 420 bases / second. Outputs may vary according to library type, run conditions, etc.

Portable

Sequence anywhere, including at sample source.

Real time

Immediate data streaming for rapid, actionable results.

Unrestricted read length

Generate short to ultra-long (>4 Mb) reads for ultimate experimental flexibility.

Zalety długich odczytów

- Umożliwia sekwencjonowanie obszarów repetytywnych
  - W 2022 pierwszy kompletny genom człowieka, łącznie z centromerami i telomerami (T2T - telomere to telomere)
  - Wykrywanie zmian strukturalnych (duże delecje, zmienna liczba kopii, rearanżacje)
- Bezpośrednie wykrywanie modyfikacji epigenetycznych (metylacja DNA) - nanopore

STATISTICS	GRCH38	T2T-CHM13	DIFFERENCE (±%)
Summary			
Assembled bases (Gbp)	2.92	3.05	+4.5
Unplaced bases (Mbp)	11.42	0	−100.0
Gap bases (Mbp)	120.31	0	−100.0
Number of contigs	949	24	−97.5
Contig NG50 (Mbp)	56.41	154.26	+173.5
Number of issues	230	46	−80.0
Issues (Mbp)	230.43	8.18	−96.5
Gene annotation			
Number of genes	60,090	63,494	+5.7
Protein coding	19,890	19,969	+0.4
Number of exclusive genes	263	3,604	
Protein coding	63	140	
Number of transcripts	228,597	233,615	+2.2
Protein coding	84,277	86,245	+2.3
Number of exclusive transcripts	1,708	6,693	
Protein coding	829	2,780	
Segmental duplications			
Percentage of segmental duplications (%)	5.00	6.61	
Segmental duplication bases (Mbp)	151.71	201.93	+33.1
Number of segmental duplications	24097	41528	+72.3
RepeatMasker			
Percentage of repeats (%)	51.89	53.94	
Repeat bases (Mbp)	1,516.37	1,647.81	+8.7
Long interspersed nuclear elements	626.33	631.64	+0.8
Short interspersed nuclear elements	386.48	390.27	+1.0
Long terminal repeats	267.52	269.91	+0.9
Satellite	76.51	150.42	+96.6

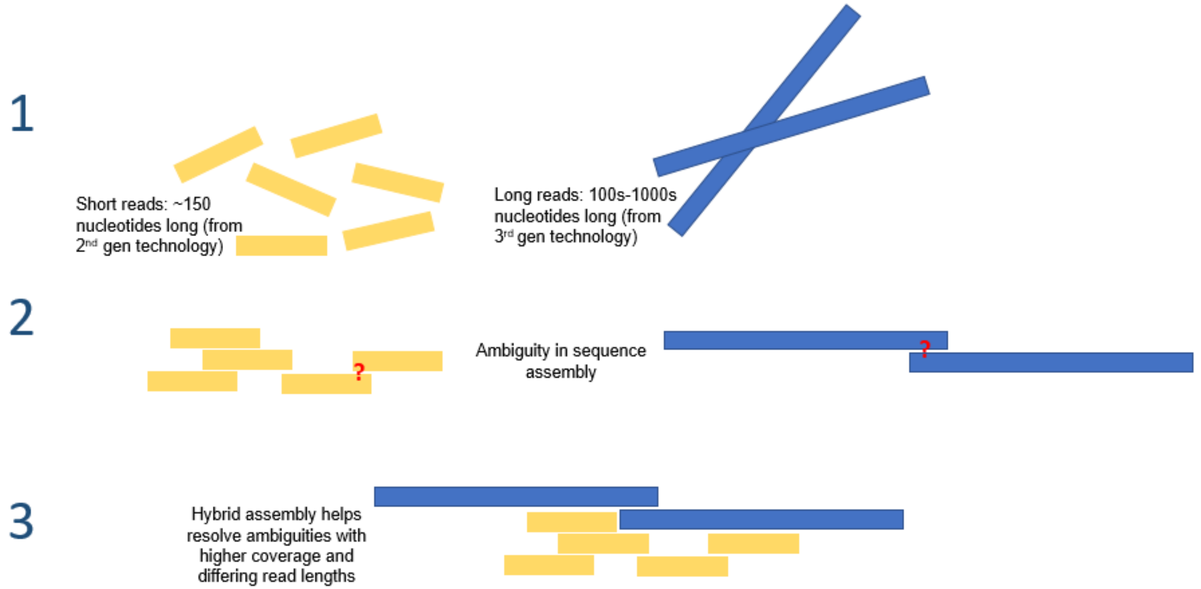
HUMAN GENOMICS

The complete sequence of a human genome

Nurk *et al.*, *Science* **376**, 44–53 (2022) 1 April 2022

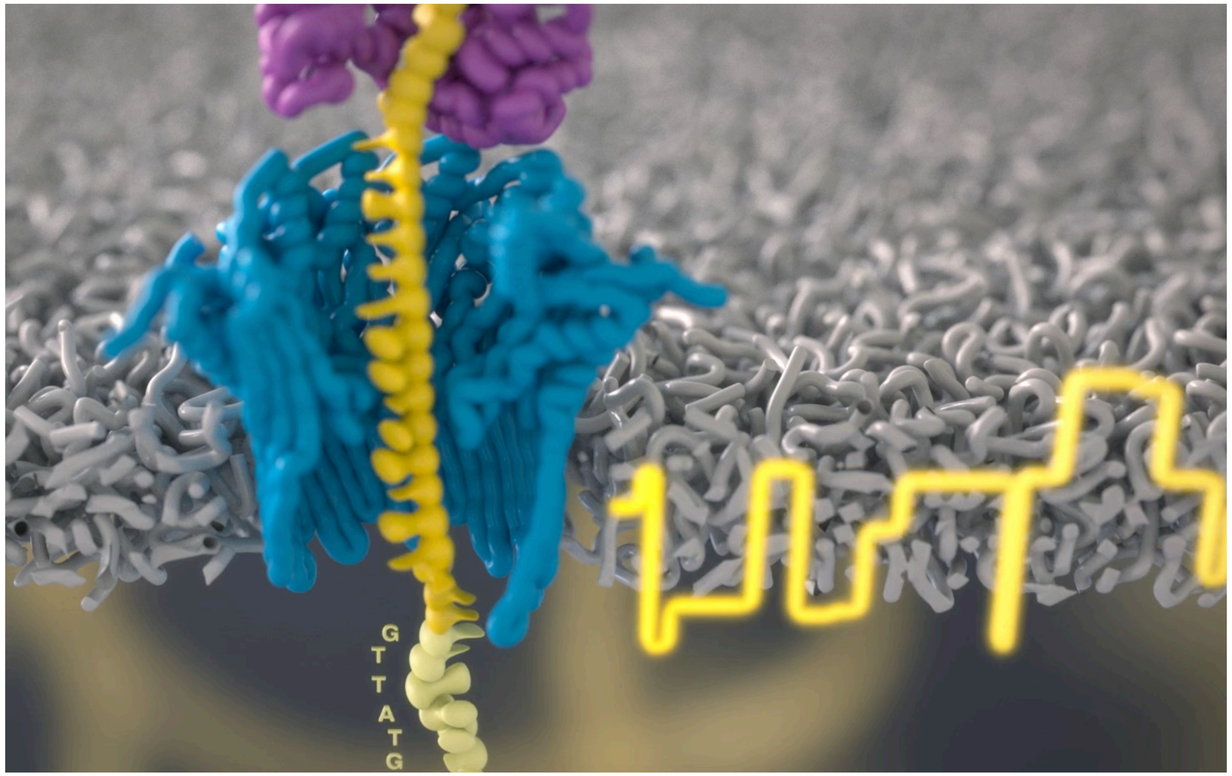
NGS trzeciej generacji - wyzwania

- Mniejsza przepustowość
- Niższa dokładność
  - szybkie postępy ostatnio
  - sekwencjonowanie hybrydowe - składanie odczytów krótkich na matrycy długich
- Narzędzia bioinformatyczne mniej rozwinięte



Ukierunkowane sekwencjonowanie nanopore

- Sekwencja odczytywana w czasie rzeczywistym
- Można usunąć lub zaakceptować na podstawie początkowej sekwencji
- Bez amplifikacji
  - zachowany stan metylacji



© nanoporetech.com

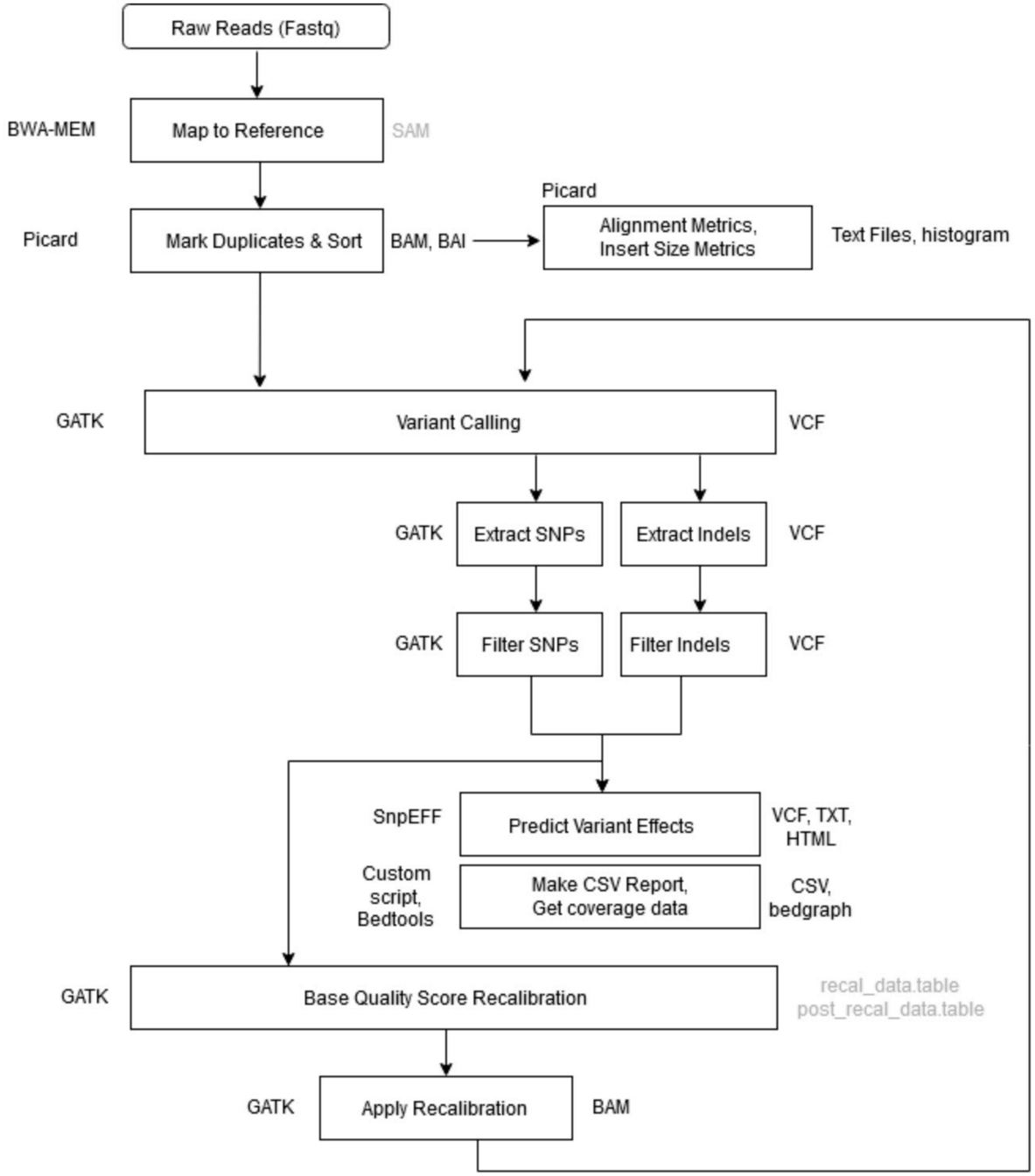
Targeted long-read sequencing identifies missing disease-causing variation

Danny E. Miller,[1,2,*] Arvis Sulovari,[1,21] Tianyun Wang,[1,21] Hailey Loucks,[2] Kendra Hoekzema,[1] Katherine M. Munson,[1] Alexandra P. Lewis,[1] Edith P. Almanza Fuerte,[2,22] Catherine R. Paschal,[3,4] Tom Walsh,[1,5] Jenny Thies,[2] James T. Bennett,[2,3,6,7] Ian Glass,[2] Katrina M. Dipple,[2,7,8] Karynne Patterson,[1] Emily S. Bonkowski,[2] Zoe Nelson,[2] Audrey Squire,[2] Megan Sikes,[2] Erika Beckman,[2] Robin L. Bennett,[5] Dawn Earl,[2] Winston Lee,[9,10] Rando Allikmets,[10,11] Seth J. Perlman,[12] Penny Chow,[13] Anne V. Hing,[13] Tara L. Wenger,[2] Margaret P. Adam,[2] Angela Sun,[2,8] Christina Lam,[2,7,14] Irene Chang,[2] Xue Zou,[15] Stephanie L. Austin,[16] Erin Huggins,[16] Alexias Safi,[16] Apoorva K. Iyengar,[17,18] Timothy E. Reddy,[17] William H. Majoros,[17] Andrew S. Allen,[17] Gregory E. Crawford,[16] Priya S. Kishnani,[16] University of Washington Center for Mendelian Genomics, Mary-Claire King,[1,5] Tim Cherry,[6] Jessica X. Chong,[2,7] Michael J. Bamshad,[1,2,7] Deborah A. Nickerson,[1,7] Heather C. Mefford,[2,7,22] Dan Doherty,[2,7,19] and Evan E. Eichler[1,7,20,*]

The American Journal of Human Genetics *108*, 1436–1449, August 5, 2021

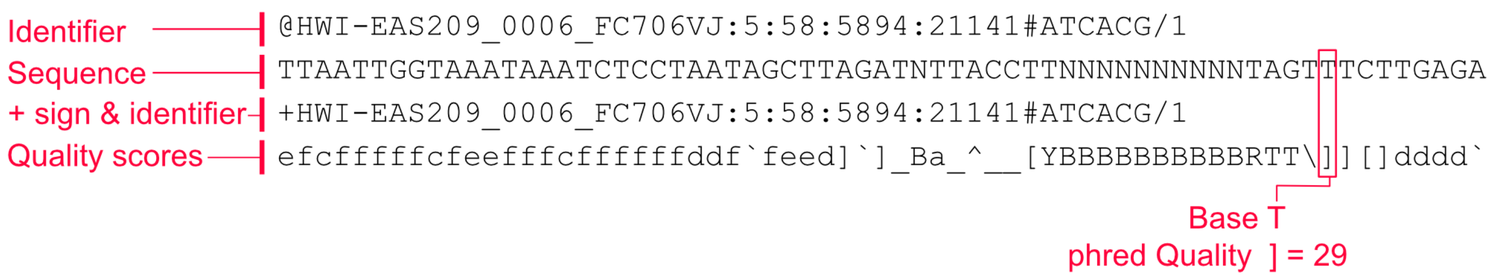
Analiza sekwencji genomu

- Mapowanie odczytów do sekwencji referencyjnej
- Identyfikacja wariantów z kontrolą jakości (!)
  - opcjonalnie: ustalenie fazy i haplotypów chromosomów
- Przewidywanie efektu i annotacja wariantów
- Filtrowanie i interpretacja wyników



Mapowanie

- Odczyty zwykle w formacie FASTQ
- Wybór genomu referencyjnego
  - obecnie z reguły GRCh38 (hg38)
- Algorytmy dobrze znane i dopracowane
  - zależnie od platformy, dla Illumina zwykle BWA-MEM
- Etap wymagający sporych zasobów komputerowych i czasu
- Wynik w formacie BAM (~50GB/genom)



https://compgenomr.github.io/book/

Identyfikacja wariantów

- Różne metody, często dające różne wyniki
  - GATK - standard
  - DeepVariant - z wykorzystaniem uczenia maszynowego
- Wyważenie pomiędzy czułością a unikaniem fałszywie pozytywnych wyników
  - wskazane filtrowanie z uwzględnieniem jakości/wiarygodności
- Etap trudny i wymagający dużych zasobów
- Nieco inne metody i parametry przy identyfikacji wariantów dziedziczonych i somatycznych
- Wynik w formacie VCF, zawiera tylko miejsca różne od genomu referencyjnego (warianty)

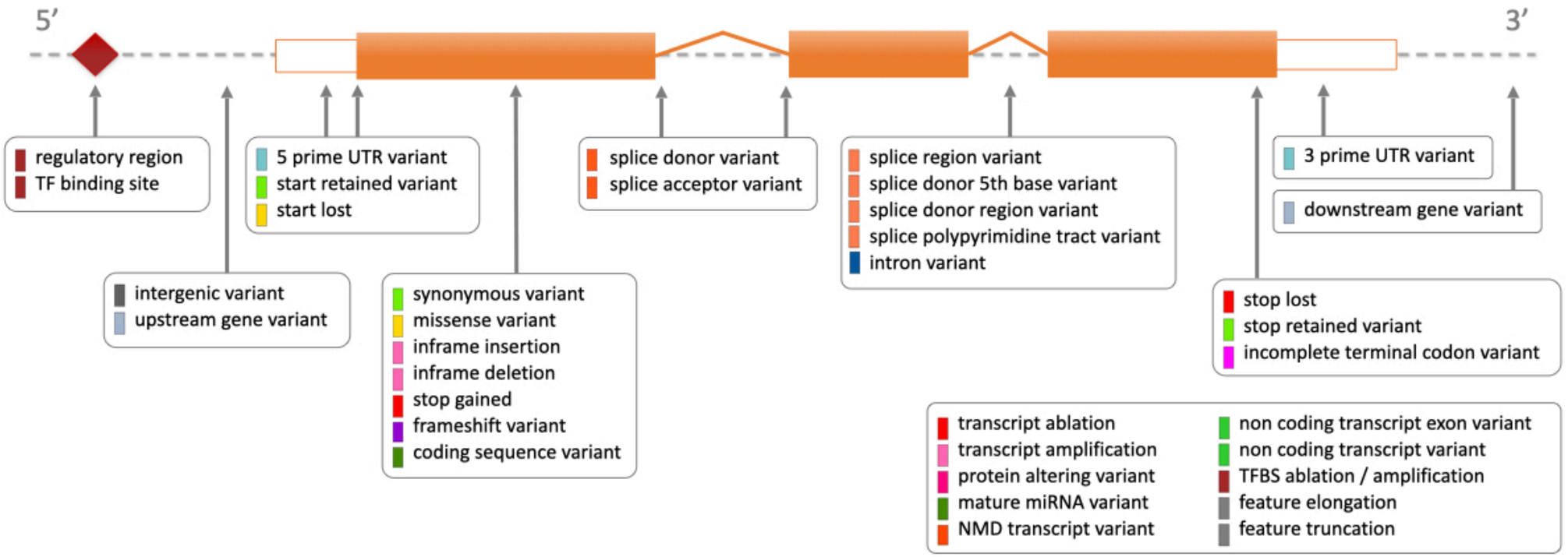
Przewidywanie efektu i annotacja wariantów

- Łączy różne rodzaje analiz
  - weryfikacja, czy dany wariant występuje w bazach danych (dbSNP, ClinVar) itp.
  - sprawdzenie, jak często dany allel występuje w dotychczas zbadanych genomach (gnomAD, 1000 genomes)
  - lokalizacja wariantu w obrębie genu (lub innego annotowanego obszaru genomu) i predykcja efektu mutacji
    - wykorzystuje standardowy zestaw terminów opisujących elementy genomu i efekty mutacji - *Sequence Ontology*
- Główne narzędzie (i źródło standardowej terminologii): ENSEMBL VEP (Variant Effect Predictor)
  - lokalnie z linii komend lub online (nie polecane do dużych zbiorów wariantów)
- Wyniki zapisane w postaci tabeli lub w dodatkowych kolumnach w pliku VCF

Bazy danych

- dbSNP - wszystkie znane krótkie warianty, identyfikator (format np. rs8176719) standardem w opisywaniu wariantów
- ClinVar - warianty o znanym efekcie klinicznym (zmienność patologiczna)
- GWAS Catalog - badania asocjacji
- SNPedia - różne związki wariantów z fenotypem na podstawie literatury, często stosowana w komercyjnych analizach DTC

Sequence Ontology



Predykcja efektów

- Szereg algorytmów przewidujących efekt mutacji na podstawie
  - konserwacji ewolucyjnej
  - przewidywanego wpływu zmiany aminokwasu na strukturę białka
  - itp.
- Najbardziej znane: SIFT, CADD, PolyPhen
- Ogólna klasyfikacja: IMPACT
  - HIGH
  - MODERATE
  - MODIFIER
  - LOW

SIFT value	Qualitative prediction	Website display example
Less than 0.05	"Deleterious"	0.01
	"Deleterious - low confidence"	0.01
Greater than or equal to 0.05	"Tolerated"	0.8
	"Tolerated - low confidence"	0.8

Polyphen value	Qualitative prediction	Website display example
greater than 0.908	"Probably Damaging"	0.95
greater than 0.446 and less than or equal to 0.908	"Possibly Damaging"	0.5
less than or equal to 0.446	"Benign"	0.25
unknown	"Unknown"	unknown

Ostateczna analiza

- Nie można przekazać wyników automatycznej annotacji odbiorcy (klinicysta, pacjent)
- Wyniki muszą być przeanalizowane przez wykwalifikowaną osobę ze wskazaniem tych wariantów, które mogą mieć znaczenie w predykcji fenotypu
  - np. znane warianty patogenne - należy sięgnąć do publikacji i ocenić wiarygodność dowodów, zweryfikować czy np. są recesywne (a czy są w genomie homozygotyczne), itp.
  - znane czynniki ryzyka - należy ocenić ryzyko biorąc pod uwagę informacje o probandzie, dane populacyjne, itp.
  - czy stwierdzono nowe, nieopisane wcześniej warianty (obecnie zdarza się rzadko, chyba że badamy rzadkie choroby)

Wielogenowe czynniki ryzyka

- GPS - *Genome-wide Polygenic Scores*/ PRS (*Polygenic Risk Scores*)
- Łączy pojedyncze asocjacje, z których każda ma nieznaczny wpływ na fenotyp
- Analiza LDpred - uwzględnia wszystkie znane SNP, waga zależna od asocjacji

SNP	Increasing allele	Allele 1	Allele 2	Genotypic score	Correlation with trait	Weighted genotypic score
SNP 1	T	A	T	1	0.005	0.005
SNP 2	C	G	G	0	0.004	0.000
SNP 3	A	A	A	2	0.003	0.006
SNP 4	G	C	G	1	0.003	0.003
SNP 5	G	C	C	0	0.003	0.000
SNP 6	T	A	T	1	0.002	0.002
SNP 7	C	C	G	1	0.002	0.002
SNP 8	A	A	A	2	0.002	0.004
SNP 9	A	T	T	0	0.001	0.000
SNP 10	C	C	G	1	0.001	0.001
Polygenic score				**9**		**0.023**

The new genetics of intelligence

Robert Plomin[1] and Sophie von Stumm[2]

Wielogenowe czynniki ryzyka

- Współczynnik ryzyka wczesnej choroby wieńcowej na podstawie analizy 182 znanych zasocjowanych polimorfizmów
- Jednoznaczna predykcja tylko dla skrajnych wartości
- To nadal nie jest diagnostyka, ale może się przydać w identyfikowaniu grup największego ryzyka

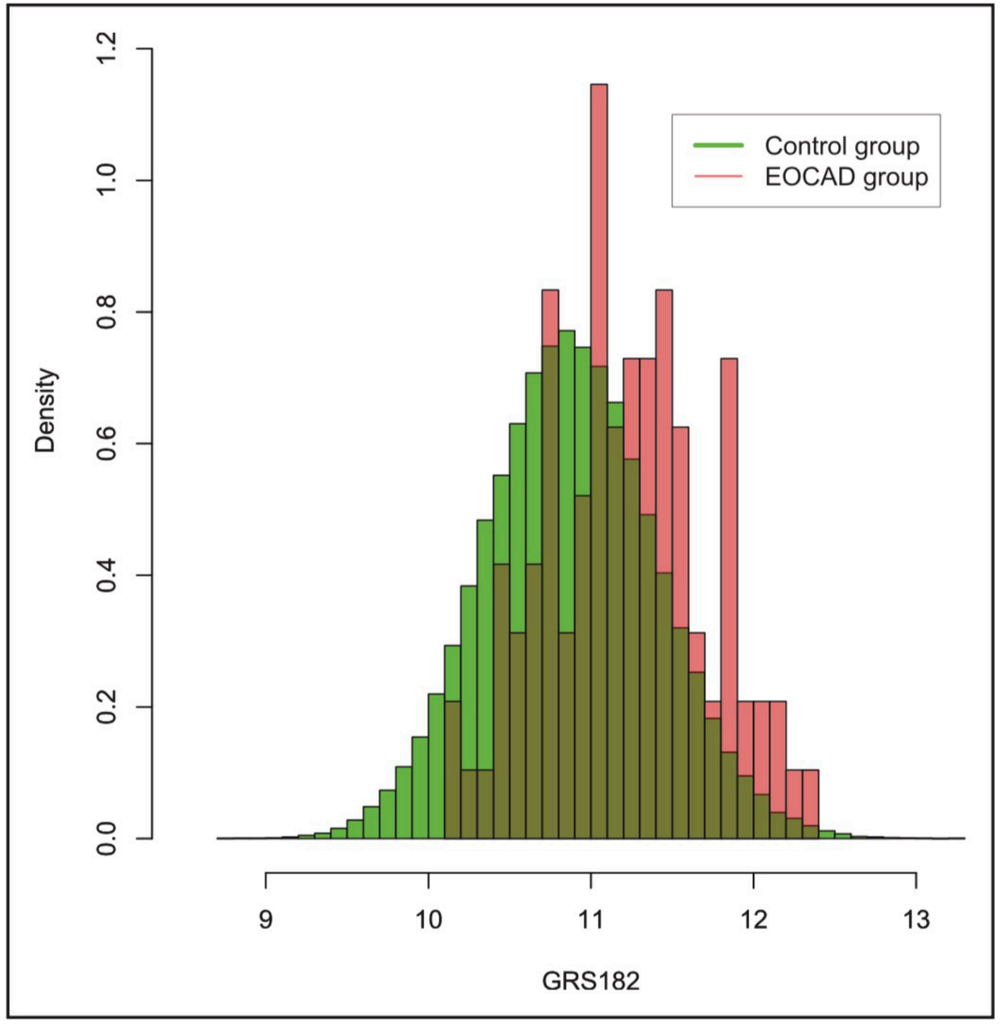


Figure 1. Distribution of genetic risk score based on the presence of 182 independent variants associated with coronary artery disease (GRS182) in the UK Biobank cohort according to early-onset coronary artery disease (EOCAD) status.

Circ Genom Precis Med. 2018;11:e001849. DOI: 10.1161/CIRCGEN.117.001849

Choroba wieńcowa

- CAD (*Coronary Arterial Disease*)
- ~60 000 chorych, ~ 120 000 kontroli (UK Biobank)
- 6 630 150 SNP

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4]*

NATURE GENETICS | VOL 50 | SEPTEMBER 2018 | 1219–1224

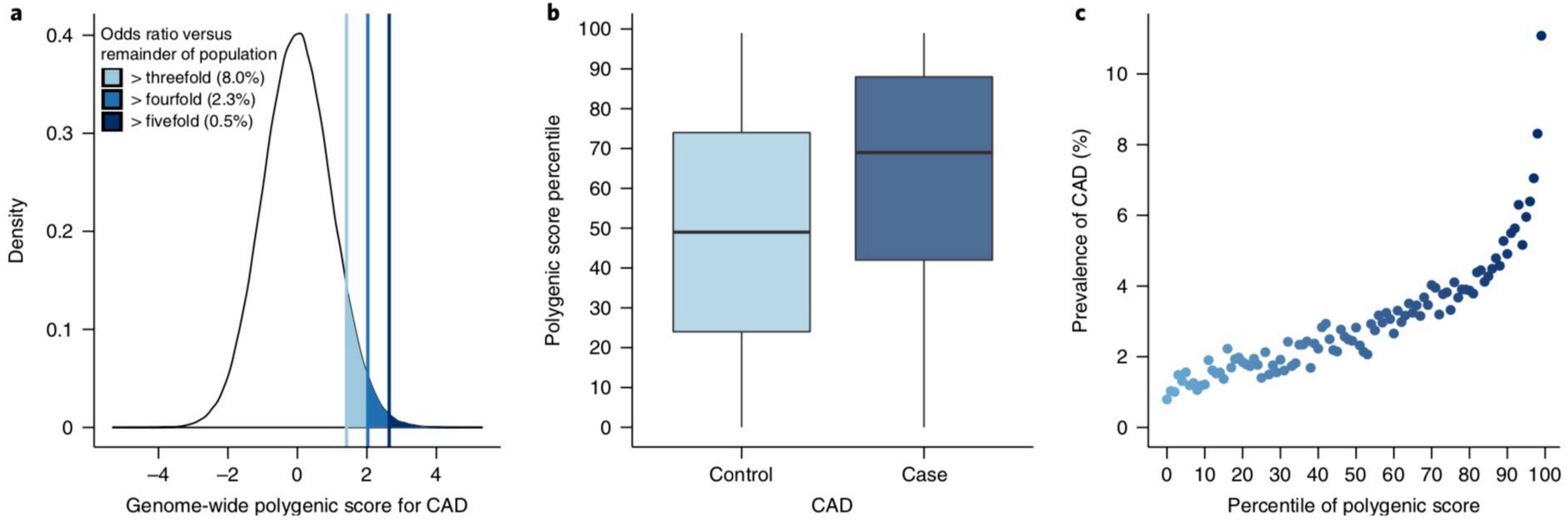


Fig. 2 | Risk for CAD according to GPS. a, Distribution of GPS_{CAD} in the UK Biobank testing dataset ($n = 288{,}978$). The x axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and

Stratyfikacja ryzyka

- Bardzo znaczny wzrost ryzyka u stosunkowo nielicznej grupy o wysokim GPS/PRS

High GPS definition	Reference group	Odds ratio	95% CI
CAD			
Top 20% of distribution	Remaining 80%	2.55	2.43–2.67
Top 10% of distribution	Remaining 90%	2.89	2.74–3.05
Top 5% of distribution	Remaining 95%	3.34	3.12–3.58
Top 1% of distribution	Remaining 99%	4.83	4.25–5.46
Top 0.5% of distribution	Remaining 99.5%	5.17	4.34–6.12

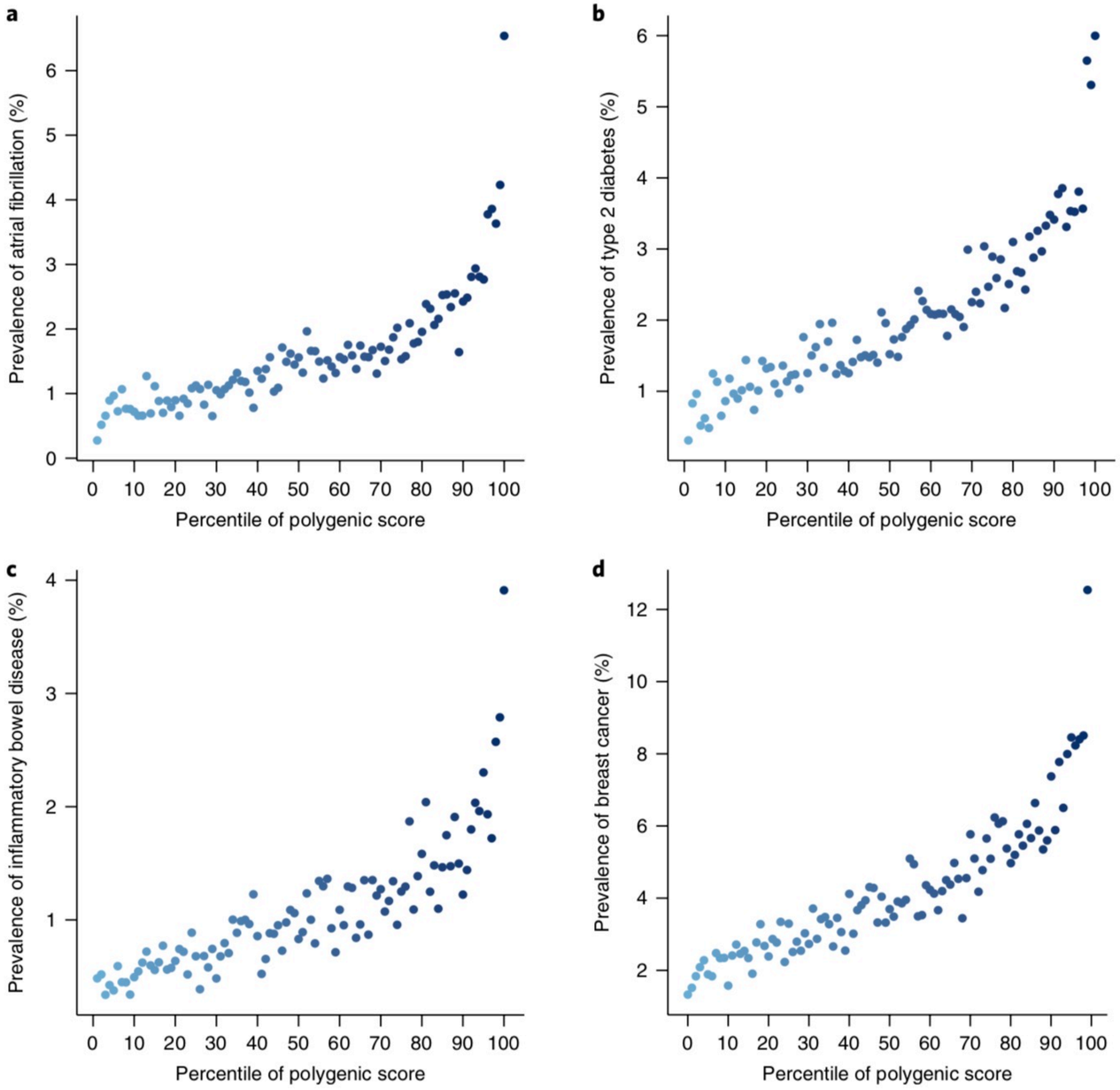
Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4]*

NATURE GENETICS | VOL 50 | SEPTEMBER 2018 | 1219–1224

Inne choroby

- Dla 4 innych badanych chorób podobny przebieg krzywej
  - (a) migotanie przedsionków, (b) cukrzyca typu 2, (c) nieswoiste zapalenie jelit, (d) rak piersi
- Dla diagnostyki i profilaktyki istotne osoby z górnych percentyli rozkładu GPS/PRS

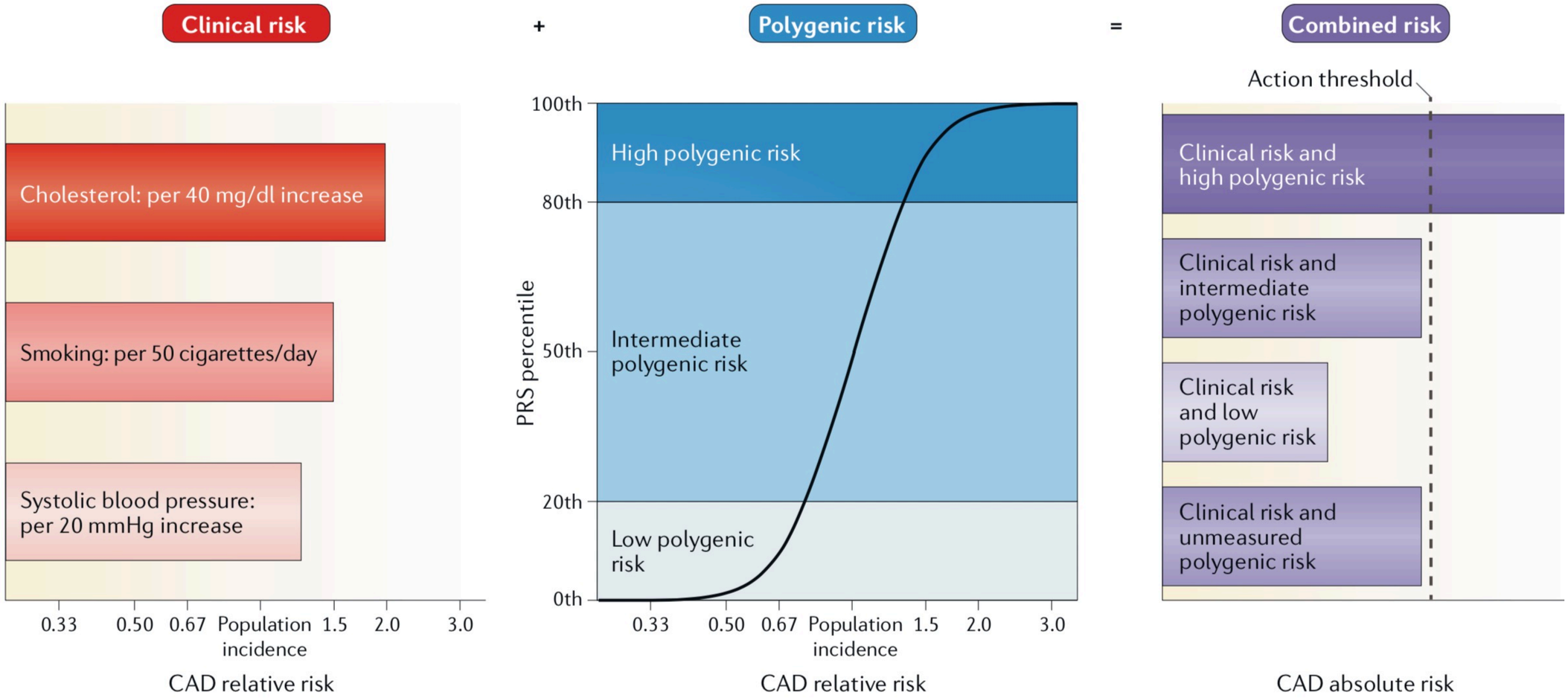


Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4*]

NATURE GENETICS | VOL 50 | SEPTEMBER 2018 | 1219–1224

Ryzyko środowiskowe i genetyczne



The personal and clinical utility of polygenic risk scores

Ali Torkamani[1,2]*, *Nathan E. Wineinger*[1,2] *and Eric J. Topol*[1,3]

Wielogenowe czynniki ryzyka

- Wyniki podawane jako
  - iloraz szans (Odds ratio, OR) - dla cech dyskretnych (np. chory/zdrowy)
  - współczynnik beta - dla cech ilościowych
- Kluczowa dla interpretacji informacja: w którym percentylu dla odpowiedniej populacji mieści się dany wynik