UNIVERSITY OF OXFORD

Bioinformatics for Genomic Medicine

Lecture – part I

Silvia Salatino, PhD

High-throughput Bioinformatician
Wellcome Centre for Human Genetics, Oxford

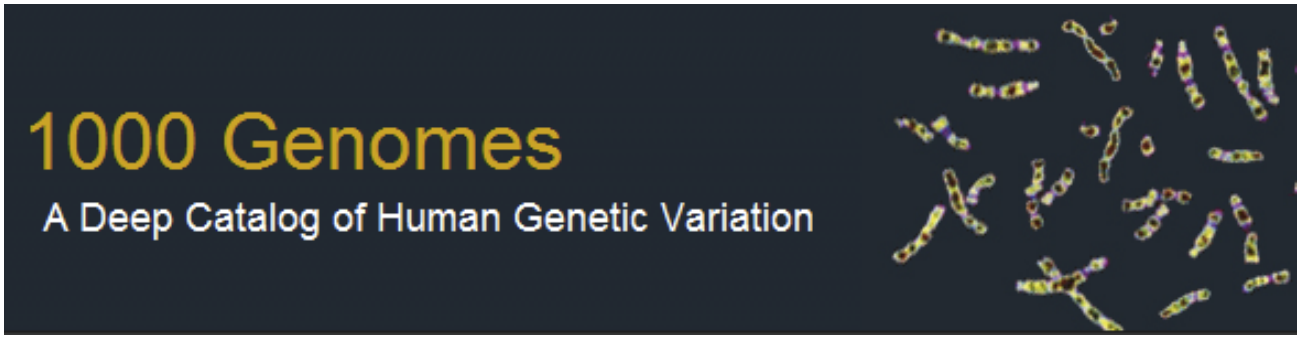
Email: silvia@well.ox.ac.uk

Module for the DPhil programme
Genomic Medicine and Statistics

22-26.10.2018

Table of contents

Major genomic projects – 1000 Genomes Project



The first project that sequenced whole genomes of a large cohort of people aiming to find most genetic variants with frequencies of at least 1% in the populations studied.

Launched at the Wellcome Genome Campus in 2007, it was completed by 2015.

>100 contributors from several institutes, corresponding Author: Gil A. McVean

Pilot	Purpose	Coverage	Strategy	Status
1 - low coverage	Assess strategy of sharing data across samples	2-4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 - trios	Assess coverage and platforms and centres	20-60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 - gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

The final data set contains data for 2'504 individuals from 26 populations. The data is free for download provided it is properly cited. The 1000 Genomes Project also developed guidelines on ethical considerations for investigators collecting the samples.

Major genomic projects – HapMap

Started in 2002, the **International HapMap Project** was a world-wide collaboration of different academic centres and private companies that aimed to develop a haplotype map (HapMap) of the human genome, to:

- describe the common patterns of human genetic variation
- find genes and genetic variations that affect health and disease
- study the genetic factors contributing to variation in response to environmental factors, in susceptibility to infection, and in the effectiveness of and adverse responses to drugs and vaccines.

Although being considered a stepping stone for the 1000 Genomes Project, it has now become obsolete and the NCBI decided to withdraw it last year:

Populations

The following population samples were studied:

ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

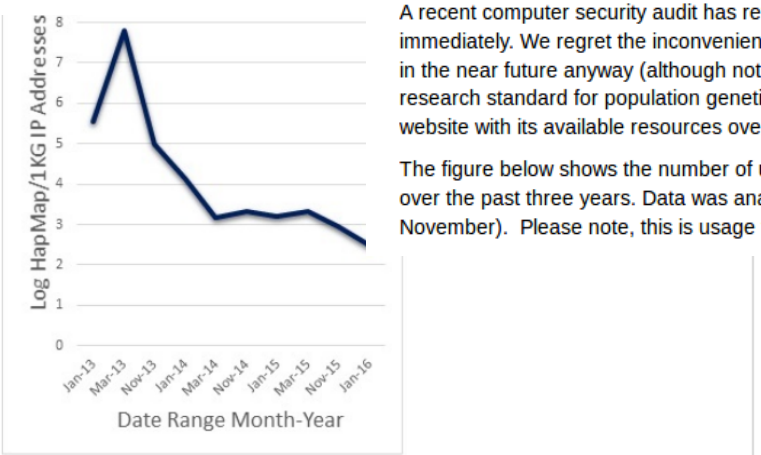


Figure 1: HapMap vs 1KG usage over the past three years exemplified during the peak usage months.

NCBI Resources How To Sign in to NCB

Variation Search NCBI Search

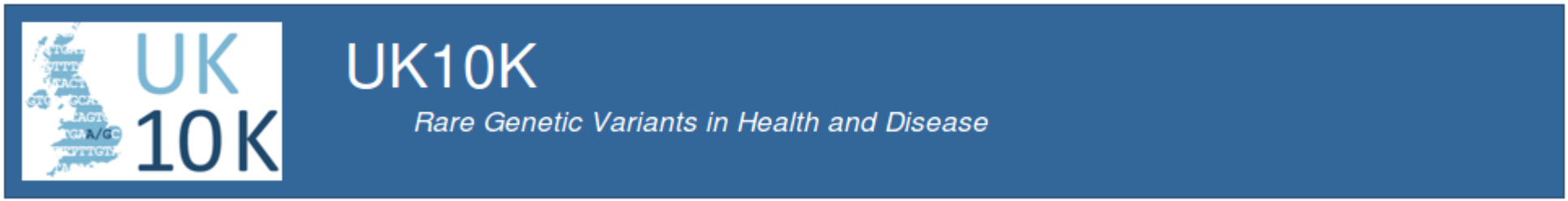
NCBI retiring HapMap Resource

June 16, 2016

A recent computer security audit has revealed security flaws in the legacy HapMap site that require NCBI to take it down immediately. We regret the inconvenience, but we are required to do this. That said, NCBI was planning to decommission this site in the near future anyway (although not quite so suddenly), as the 1,000 genomes (1KG) project has established itself as a research standard for population genetics and genomics. NCBI has observed a decline in usage of the HapMap dataset and website with its available resources over the past five years and it has come to the end of its useful life.

The figure below shows the number of unique IP addressing accessing HapMap relative to the 1KG website has been declining over the past three years. Data was analyzed over three week sections for the peak usage months (January, March and November). Please note, this is usage for NCBI only, and many users access 1KG data from EBI.

Major genomic projects – UK10K



The UK10K project aimed to help uncover rare variants contributing to disease by studying the genetic code of 10'000 people at an order of magnitude deeper than the 1000 Genomes Project. In particular, the consortium performed:

- **Whole genome sequencing of 4'000 people** whose physical characteristics are well documented, trying to identify those changes that have no discernible effect and those that may be linked to a particular disease

- **Exome sequencing of 6'000 people** with extreme health problems, in order to study the changes within protein-coding areas of DNA that tell the body how to make proteins and comparing them with the first group

The project started in 2010 to provide a genotype/phenotype open access resource to all researchers.

Major genomic projects – The 100'000 Genomes Project

Genomics england

Google Custom Search Search

About Us | 100,000 Genomes Project | Taking Part | For Healthcare Professionals | Research | Industry Partnerships | News & Events

Home › The 100,000 Genomes Project

The 100,000 Genomes Project

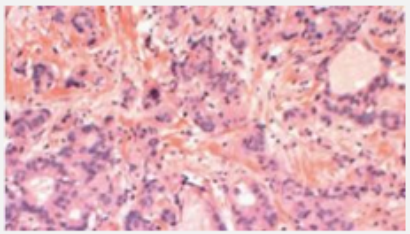
The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

The project will also enable new medical research. Combining genomic sequence data with medical records is a ground-breaking resource. Researchers will study how best to use genomics in healthcare and how best to interpret the data to help patients. The causes, diagnosis and treatment of disease will also be investigated. We also aim to kick-start a UK genomics industry. This is currently the largest national sequencing project of its kind in the world.

Useful links

Cancer
Introduction to cancer in the 100,000 Genomes Project.



Major genomic projects – WGS500

The Wellcome Trust Centre for Human Genetics

Home | About us | Research | News | Work and study | Contact | Internal

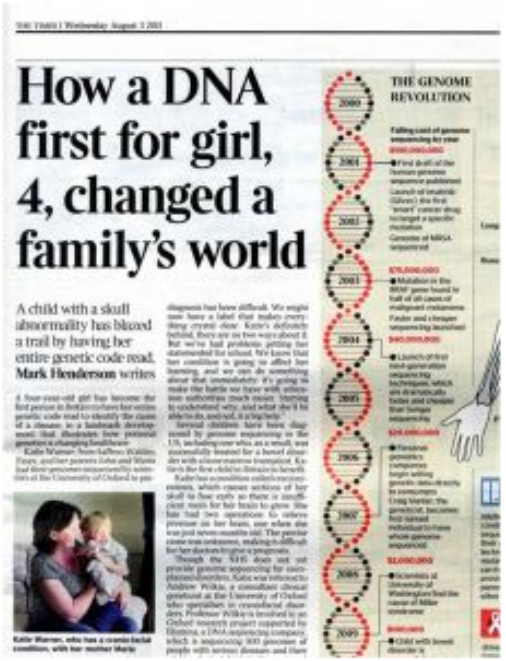
Research areas | Research groups | Cores | Publications | Podcasts: Meet our researchers | Seminars | High profile sem

Taylor group

WGS500

WGS500 is a collaborative project between the Wellcome Trust Centre for Human Genetics, the BRC Genomic Medicine Theme, and the technology company Illumina with the aim of evaluating the clinical utility of whole genome sequencing across a number of human diseases. Proposals were invited from clinicians for cases where standard genetic tests had proved negative or where no test was available. The genomes of five hundred patients and family members were sequenced spanning a range of diseases including Mendelian disorders, severe and early onset immunological conditions, and cancer, with the hope of identifying variants in novel genes or pathways to inform diagnosis, prognosis, and reproductive risk, or influence treatment selection. By mid-2013, this project had resulted in the conclusive identification of several new causative genes, with many more in validation, leading to the improvement of diagnostic genetic tests for inherited diseases in the NHS, and clearly demonstrating the value of this approach.

Publications

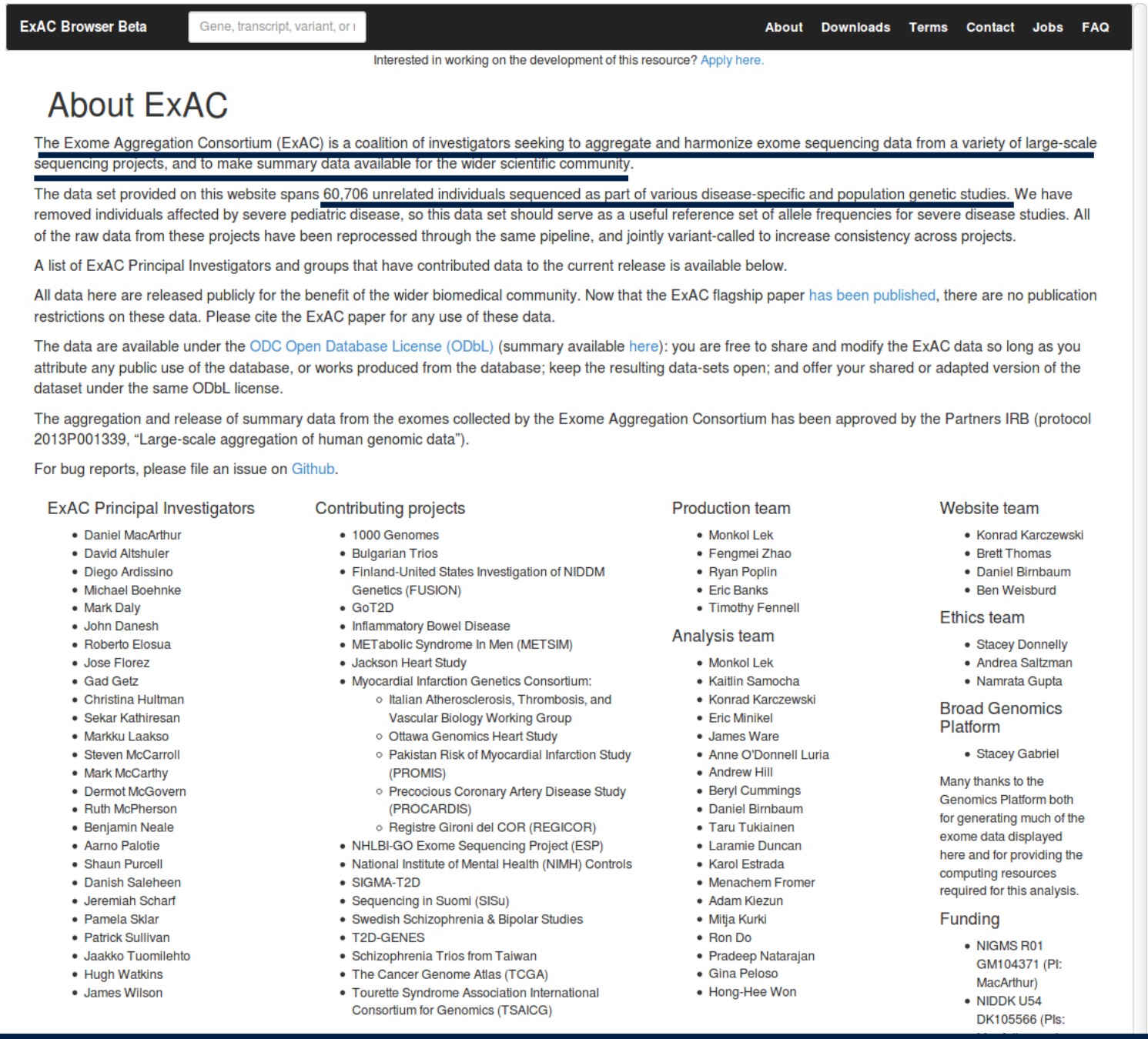


How a DNA first for girl, 4, changed a family's world

A child with a skull abnormality has blazed a trail by having her entire genetic code read. Mark Henderson writes

THE GENOME REVOLUTION

Major genomic projects – ExAC



ExAC Browser Beta | Gene, transcript, variant, or r | About | Downloads | Terms | Contact | Jobs | FAQ

Interested in working on the development of this resource? Apply here.

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. We have removed individuals affected by severe pediatric disease, so this data set should serve as a useful reference set of allele frequencies for severe disease studies. All of the raw data from these projects have been reprocessed through the same pipeline, and jointly variant-called to increase consistency across projects.

A list of ExAC Principal Investigators and groups that have contributed data to the current release is available below.

All data here are released publicly for the benefit of the wider biomedical community. Now that the ExAC flagship paper has been published, there are no publication restrictions on these data. Please cite the ExAC paper for any use of these data.

The data are available under the ODC Open Database License (ODbL) (summary available here): you are free to share and modify the ExAC data so long as you attribute any public use of the database, or works produced from the database; keep the resulting data-sets open; and offer your shared or adapted version of the dataset under the same ODbL license.

The aggregation and release of summary data from the exomes collected by the Exome Aggregation Consortium has been approved by the Partners IRB (protocol 2013P001339, "Large-scale aggregation of human genomic data").

For bug reports, please file an issue on Github.

ExAC Principal Investigators

- Daniel MacArthur
- David Altshuler
- Diego Ardissino
- Michael Boehnke
- Mark Daly
- John Danesh
- Roberto Elosua
- Jose Florez
- Gad Getz
- Christina Hultman
- Sekar Kathiresan
- Markku Laakso
- Steven McCarroll
- Mark McCarthy
- Dermot McGovern
- Ruth McPherson
- Benjamin Neale
- Aarno Palotie
- Shaun Purcell
- Danish Saleheen
- Jeremiah Scharf
- Pamela Sklar
- Patrick Sullivan
- Jaakko Tuomilehto
- Hugh Watkins
- James Wilson

Contributing projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
  - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
  - Ottawa Genomics Heart Study
  - Pakistan Risk of Myocardial Infarction Study (PROMIS)
  - Precocious Coronary Artery Disease Study (PROCARDIS)
  - Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP)
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

Production team

- Monkol Lek
- Fengmei Zhao
- Ryan Poplin
- Eric Banks
- Timothy Fennell

Analysis team

- Monkol Lek
- Kaitlin Samocha
- Konrad Karczewski
- Eric Minikel
- James Ware
- Anne O'Donnell Luria
- Andrew Hill
- Beryl Cummings
- Daniel Birnbaum
- Taru Tukiainen
- Laramie Duncan
- Karol Estrada
- Menachem Fromer
- Adam Kiezun
- Mitja Kurki
- Ron Do
- Pradeep Natarajan
- Gina Peloso
- Hong-Hee Won

Website team

- Konrad Karczewski
- Brett Thomas
- Daniel Birnbaum
- Ben Weisburd

Ethics team

- Stacey Donnelly
- Andrea Saltzman
- Namrata Gupta

Broad Genomics Platform

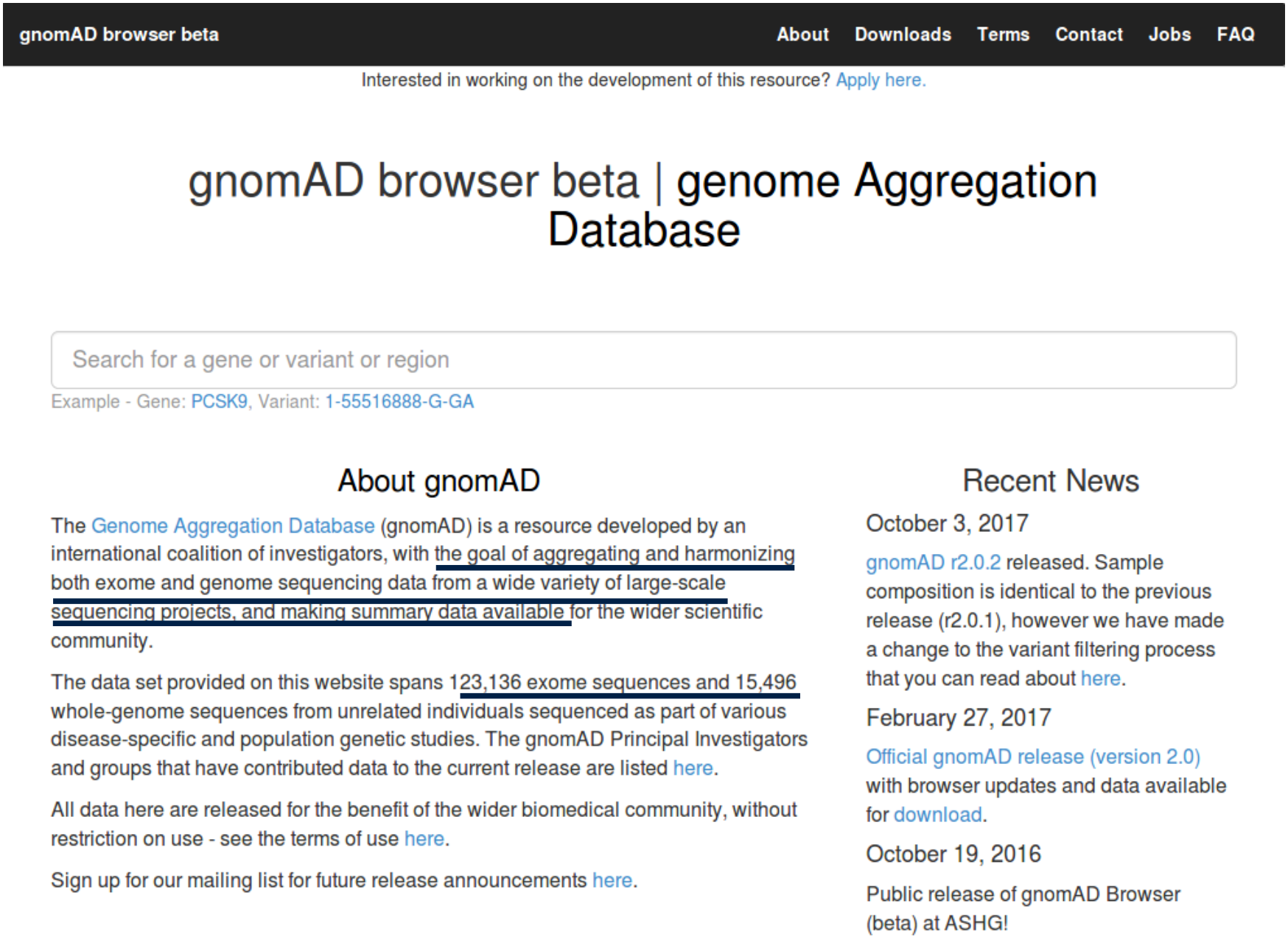
- Stacey Gabriel

Many thanks to the Genomics Platform both for generating much of the exome data displayed here and for providing the computing resources required for this analysis.

Funding

- NIGMS R01 GM104371 (PI: MacArthur)
- NIDDK U54 DK105566 (PIs:

Major genomic projects – gnomAD



Major genomic projects – POBI

People of the British Isles

UNIVERSITY OF OXFORD

Study Areas | Media | Feedback | FAQ | Volunteer Information | E-Volunteering | Login | Register

for full link, click here please!

Latest Newsletter March 2015

First paper of "People of British Isles" project

Commentary on the paper

Newsletter 5

Blog

Welcome to People of the British Isles

BREAKING NEWS : THIS IS OUR LATEST NEWSLETTER FROM MARCH 2015

Did you know that historical patterns of people's movements, from Anglo-Saxon invasions to those of the Vikings and Normans, may have an impact on 21st Century medical science?

This project began in 2004 when we were given funding by the Wellcome Trust to collect blood samples from 4,500 people from rural populations throughout the British Isles. These are being used to look at the patterns of differences in people's genetic make up around the UK. The project has two purposes, the first to help medical research, and the second to shed light on ancient migrations within the British Isles.

As part of this study, we are also interested in the inherited variation of facial features. Further funding from the Wellcome Trust has been given to return to our volunteers and collect 3D photographs of their faces, with the aim of identifying genes behind particular facial features.

Background

Your genetics differs from that of your neighbour and this means you may differ in your risk of getting particular diseases....

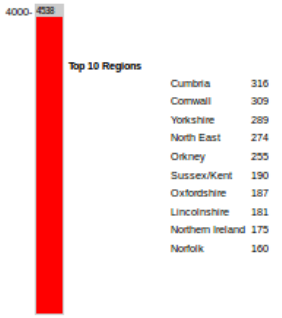
Practicalities

We have collected blood samples from about 4,500 volunteers from different rural regions throughout the UK....

First results

Our first paper with some pilot results came out online in August 2011 and is now out in printed form...

Collection Progress



How to Contact Us

Click on the map for contact details

Forthcoming Events in 2017

There will be face photographs coming on the following days, for further details please contact the co-ordinators by clicking e icon.

Friday 17th April 2015 from 10.30 am to 5.00 p.m

St Peter's Church Hall, High West Street, Dorchester, Dorset DT1 1XA

Volunteers who wish to attend and don't have an appointment can just

Major genomic projects – TCGA (ICGC)

THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Home | About Cancer Genomics | Cancers Selected for Study | Research Highlights | Publicati

Home > About TCGA > Program Overview

Program Overview

There are at least 200 forms of cancer, and many more subtypes. Each of these is caused by errors in DNA that cause cells to grow uncontrolled. Identifying the changes in each cancer's complete set of DNA – its genome – and understanding how such changes interact to drive the disease will lay the foundation for improving cancer prevention, early detection and treatment.

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset, 2.5 petabytes of data describing tumor tissue and matched normal tissues from more than 11,000 patients, is publically available and has been used widely by the research community. The data have contributed to more than a thousand studies of cancer by independent researchers and to the TCGA research network publications.

TCGA created a genomic data analysis pipeline that can effectively collect, select, and analyze human tissues for genomic alterations on a very large scale. The success of this national network of research and technology teams serves as a model for future projects and exemplifies the tremendous power of teamwork in science.

Though TCGA is coming to a close in 2017, new NCI genomics initiatives, run through the NCI Center for Cancer Genomics (CCG), will continue to build upon the success of TCGA by using the same model of collaboration for large-scale genomic analysis and by making the genomics data publically available.

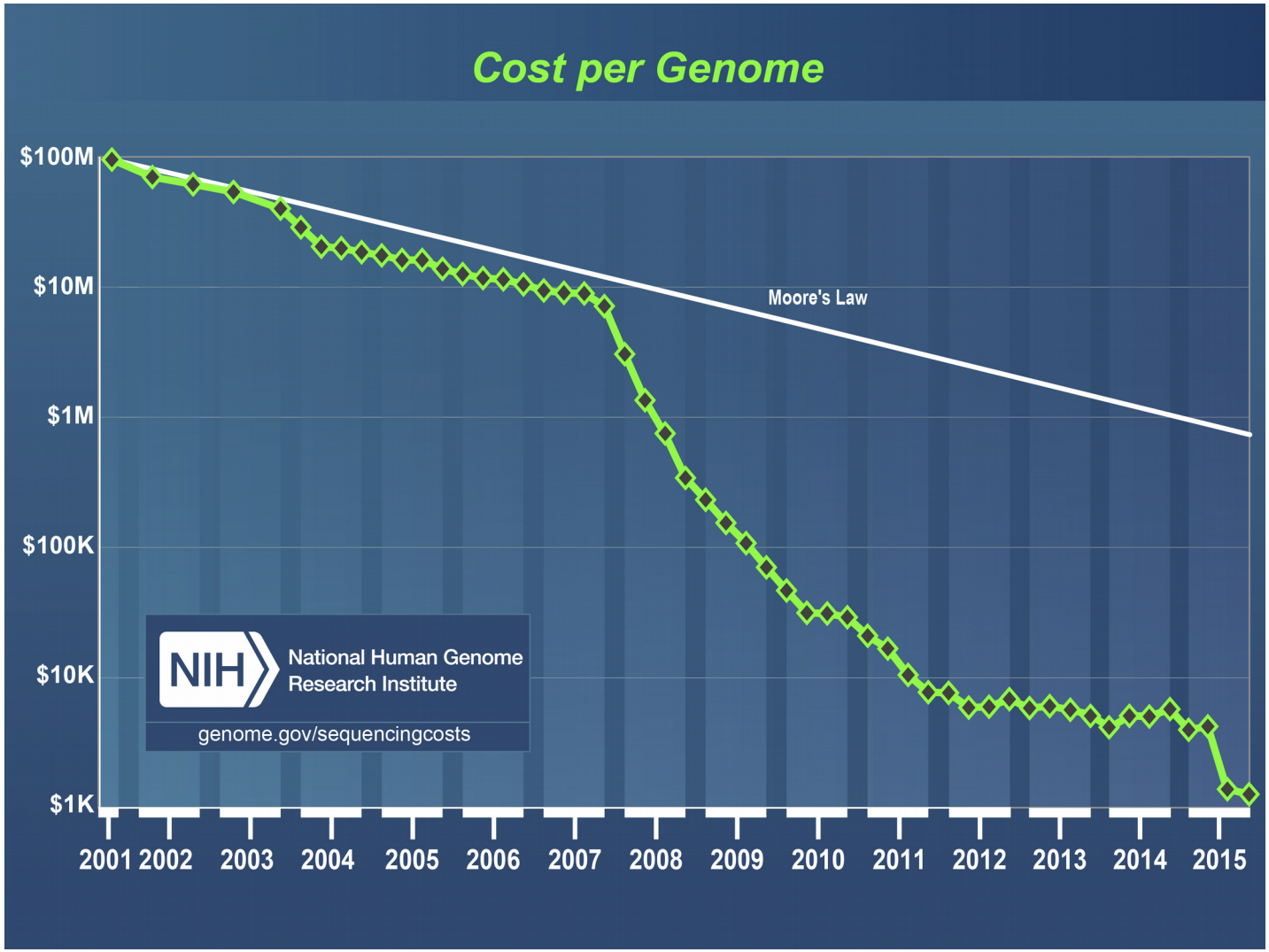
TCGA is the biggest component of the International Cancer Genome Consortium (ICGC), a collaboration of scientists from 16 nations that has discovered nearly 10 million cancer-related mutations.

Recent advances in sequencing technology

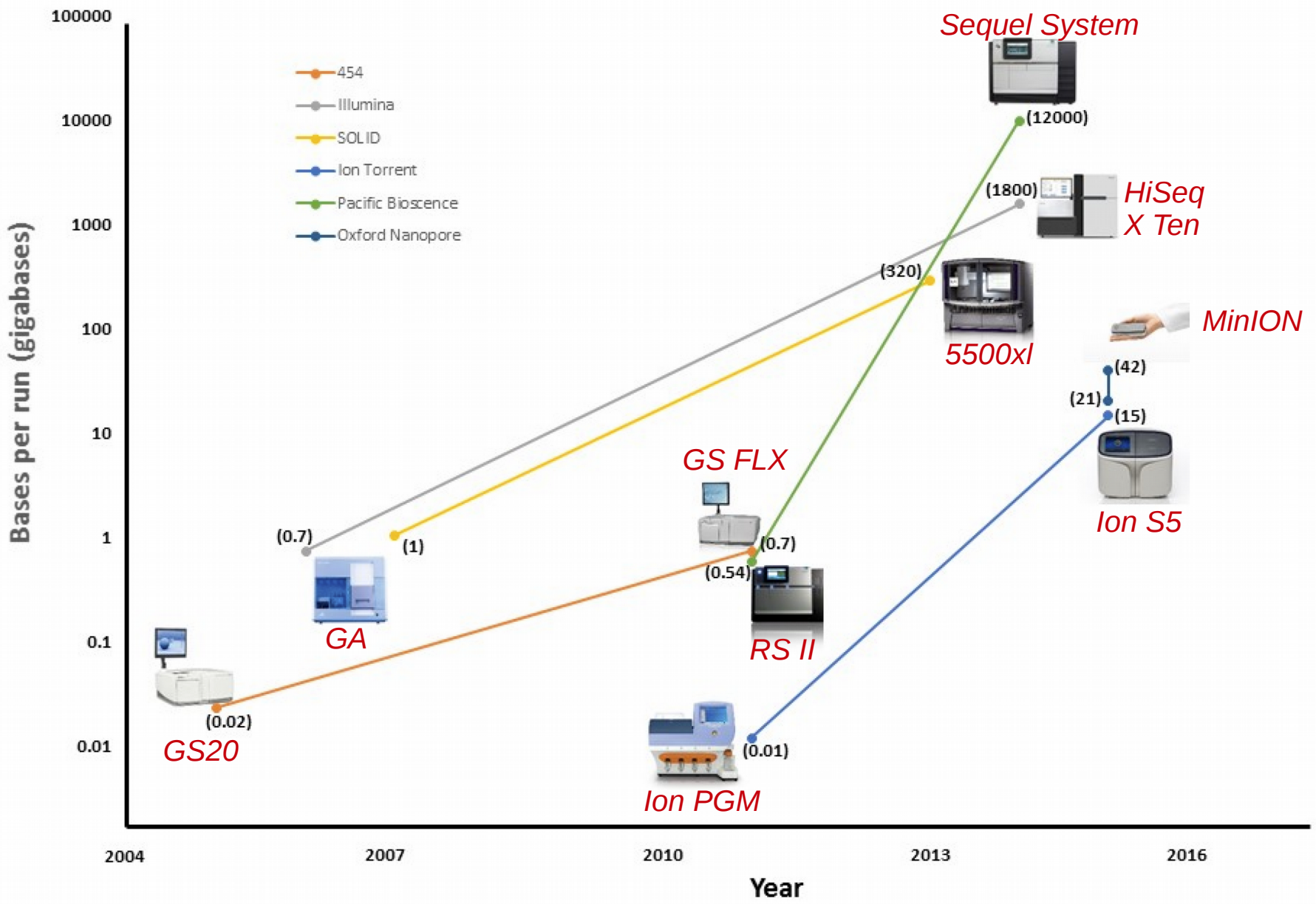
The sequencing cost has significantly dropped during the last decade, with a 10'000-fold reduction that made it possible to reach the target of sequencing a human genome for USD $ 1'000 (announced by Illumina in 2014, through the HiSeq X Ten machine), thus marking the beginning of a new era of personalized medicine.

Sequencing costs, particularly after 2008, did not decrease proportionally to the capacity of the hardware used for sequencing, as stated by Moore's Law.

This was due, mostly, thanks to the introduction of new generation sequencing (NGS) techniques, a revolution compared to the traditional Sanger sequencing.



The timeline of DNA reading capacity per platform



Garrido-Cardenas et al. Sensors 2017 Mar 14;17(3). pii: E588.

Technology targets *(adapted from K. Kassahn's slides)*

Reduction of

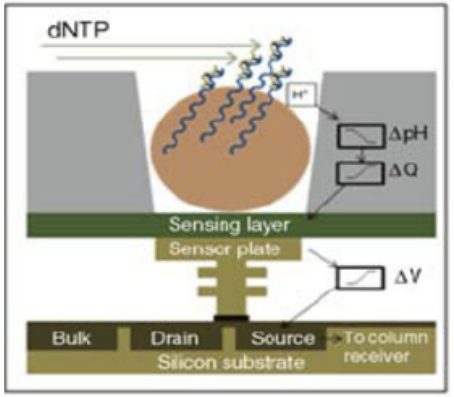
- Sequencing cost
- DNA input material
- Workflow duration
- Platform size

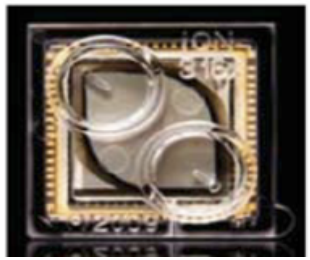
Increase of

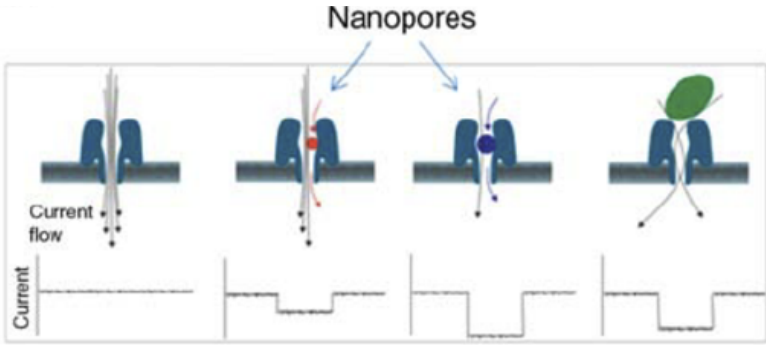
- Sequencing accuracy
- Read length
- Detection of DNA base modifications
- Single-molecule sequencing (→ less PCR cycles)

Key concepts:

- **miniaturisation**
- **parallelisation**







Genetic variation – types of mutations

The nucleus of a human somatic cell contains:

- 22 autosomal pairs + 1 pair of sex chromosomes (XX or XY)
- One set of chromosomes inherited from each parent
- Multiple copies of mitochondrial circular DNA from the mother

Germline mutation

- Inherited from the parents
- Present in every cell of the organism
- If a mutant sex cell participates in fertilization, the mutation is passed to the offspring

Somatic mutation

- Not inherited from the parents
- Present only in a portion of the organism
- Acquired from spontaneous mutations during DNA replication (mitosis), environmental factors (e.g. chemicals, UV or X-rays)
- Can result in cancer

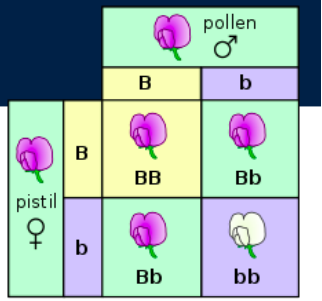
Polymorphisms

- ***SNP***: Single Nucleotide Polymorphism, found at a frequency of 1% or higher in the population (1 in every 300 bp, ~10 millions in the human genome)
- ***CNP***: Copy Number Polymorphism, structural change in the # of copies of a part of a chromosome

Types of variants

- ***Indel***: small (1bp – 1kb) insertion or deletion
- ***SNP***: 1bp mutation (does not cause a frameshift)
- ***Structural variant***: large (> 1kb) variant; e.g.: inversions, translocations, copy number variants (CNVs)

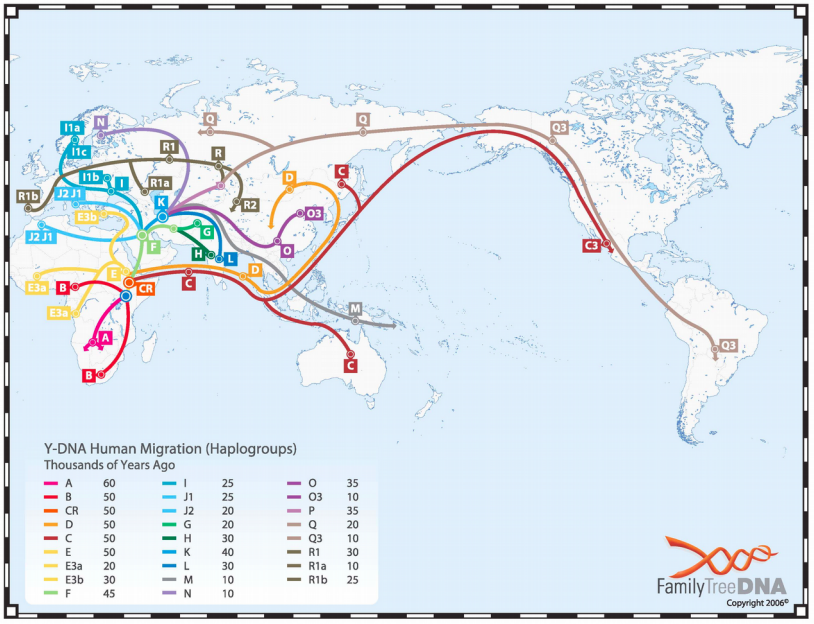
Genetic variation – genotype and haplotype

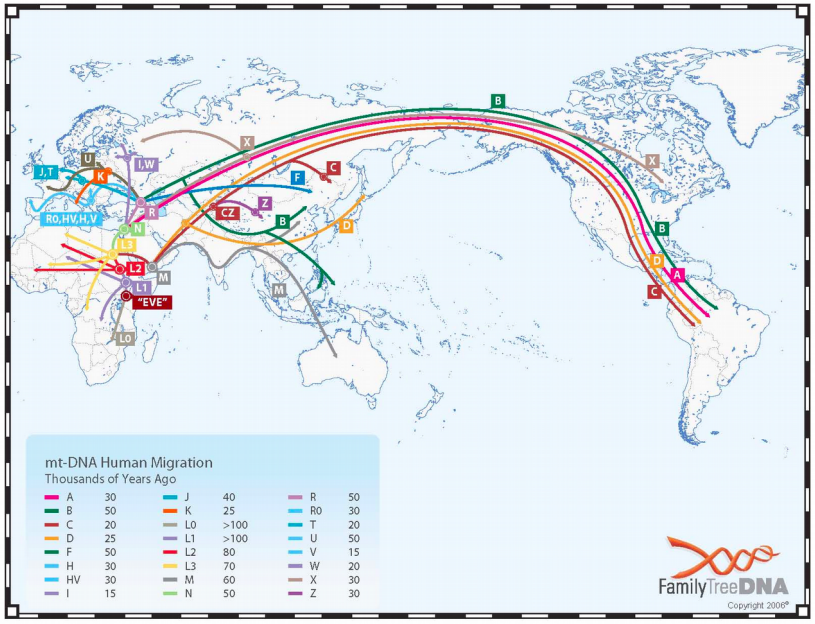


Genotype: single or multiple traits inherited together from the parents at a particular locus. The genotype determines a specific characteristic (**phenotype**) of a given cell / organism / individual, and is one of three factors that determine the phenotype (the other two being inherited epigenetic factors, and non-inherited environmental factors).

Haplotype: set of DNA variations (SNPs or STRs) that are usually inherited together from a single parent, being located on the same chromosome. The alleles making up a haplotype can be located in different places on the chromosome.

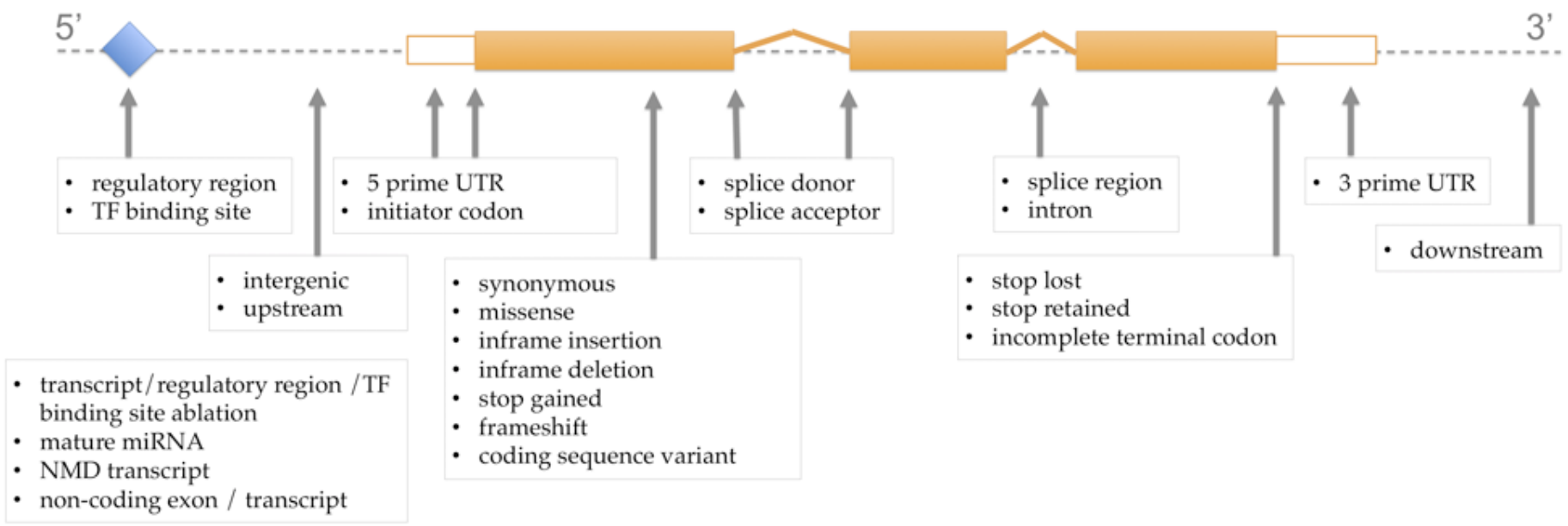
Haplogroup: the SNPs that represent the clade (= set of people sharing a common ancestor) to which a collection of particular human haplotypes belong. The most commonly studied ones are Y-chromosome (Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations and human migrations.





Genetic variation – downstream effects

Variants can have different effects on the adjacent or overlapping transcripts.



Ensembl classifies variants based on their impact in four categories:

HIGH: *The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.*

MODERATE: *A non-disruptive variant that might change protein effectiveness.*

LOW: *Assumed to be mostly harmless or unlikely to change protein behaviour.*

MODIFIER: *Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.*

Incidental findings (1)

ACMG POLICY STATEMENT | Genetics inMedicine

ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing

Robert C. Green, MD, MPH[1,2], Jonathan S. Berg, MD, PhD[3], Wayne W. Grody, MD, PhD[4–6], Sarah S. Kalia, ScM, CGC[1], Bruce R. Korf, MD, PhD[7], Christa L. Martin, PhD, FACMG[8], Amy L. McGuire, JD, PhD[9], Robert L. Nussbaum, MD[10], Julianne M. O'Daniel, MS, CGC[3], Kelly E. Ormond, MS, CGC[11], Heidi L. Rehm, PhD, FACMG[2,12], Michael S. Watson, PhD, FACMG[13], Marc S. Williams, MD, FACMG[14] and Leslie G. Biesecker, MD[15]

Disclaimer: These recommendations are designed primarily as an educational resource for medical geneticists and other health-care providers to help them provide quality medical genetic services. Adherence to these recommendations does not necessarily ensure a successful medical outcome. These recommendations should not be considered inclusive of all proper procedures and tests or exclusive of other procedures and tests that are reasonably directed to obtaining the same results. In determining the propriety of any specific procedure or test, geneticists and other clinicians should apply their own professional judgment to the specific clinical circumstances presented by the individual patient or specimen. It may be prudent, however, to document in the patient's record the rationale for any significant deviation from these recommendations.

*"In clinical exome and genome sequencing, there is a potential for the recognition and reporting of **incidental or secondary findings** unrelated to the indication for ordering the sequencing but of medical value for patient care."*

Incidental findings (2)

Primary finding = "pathogenic alterations in a gene relevant to the diagnostic indication for which the sequencing was ordered."

Incidental finding = unexpected positive (secondary) findings, namely "pathogenic or likely pathogenic alterations in genes that are not apparently relevant to a diagnostic indication for which the sequencing test was ordered."

Table 1 Conditions, genes, and variants recommended for return of incidental findings in clinical sequencing

Phenotype	MIM-disorder	PMID-Gene Reviews entry	Typical age of onset	Gene	MIM-gene	Inheritance[a]	Variants to report[b]
Hereditary breast and ovarian cancer	604370 612555	20301425	Adult	*BRCA1* *BRCA2*	113705 600185	AD	KP and EP
Li–Fraumeni syndrome	151623	20301488	Child/adult	*TP53*	191170	AD	KP and EP
Peutz–Jeghers syndrome	175200	20301443	Child/adult	*STK11*	602216	AD	KP and EP
Lynch syndrome	120435	20301390	Adult	*MLH1* *MSH2* *MSH6* *PMS2*	120436 609309 600678 600259	AD	KP and EP
Familial adenomatous polyposis	175100	20301519	Child/adult	*APC*	611731	AD	KP and EP
MYH-associated polyposis; adenomas, multiple colorectal, *FAP* type 2; colorectal adenomatous polyposis, autosomal recessive, with pilomatricomas	608456 132600	23035301	Adult	*MUTYH*	604933	AR[c]	KP and EP
Von Hippel–Lindau syndrome	193300	20301636	Child/adult	*VHL*	608537	AD	KP and EP
Multiple endocrine neoplasia type 1	131100	20301710	Child/adult	*MEN1*	613733	AD	KP and EP
Multiple endocrine neoplasia type 2	171400 162300	20301434	Child/adult	*RET*	164761	AD	KP
(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)

NOTE: KP means known pathogenic, EP means expected pathogenic

RECOMMENDATIONS

1. Constitutional mutations found in the genes on the minimum list (Table 1) should be reported by the laboratory to the ordering clinician, regardless of the indication for which the clinical sequencing was ordered.
   - Additional genes may be analyzed for incidental variants, as deemed appropriate by the laboratory.
   - Incidental variants should be reported regardless of the age of the patient.

(…)

- For most genes, only variants that have been previously reported and are a recognized cause of the disorder or variants that are previously unreported but are of the type that is expected to cause the disorder, as defined by prior ACMG guidelines,[20] should be reported.

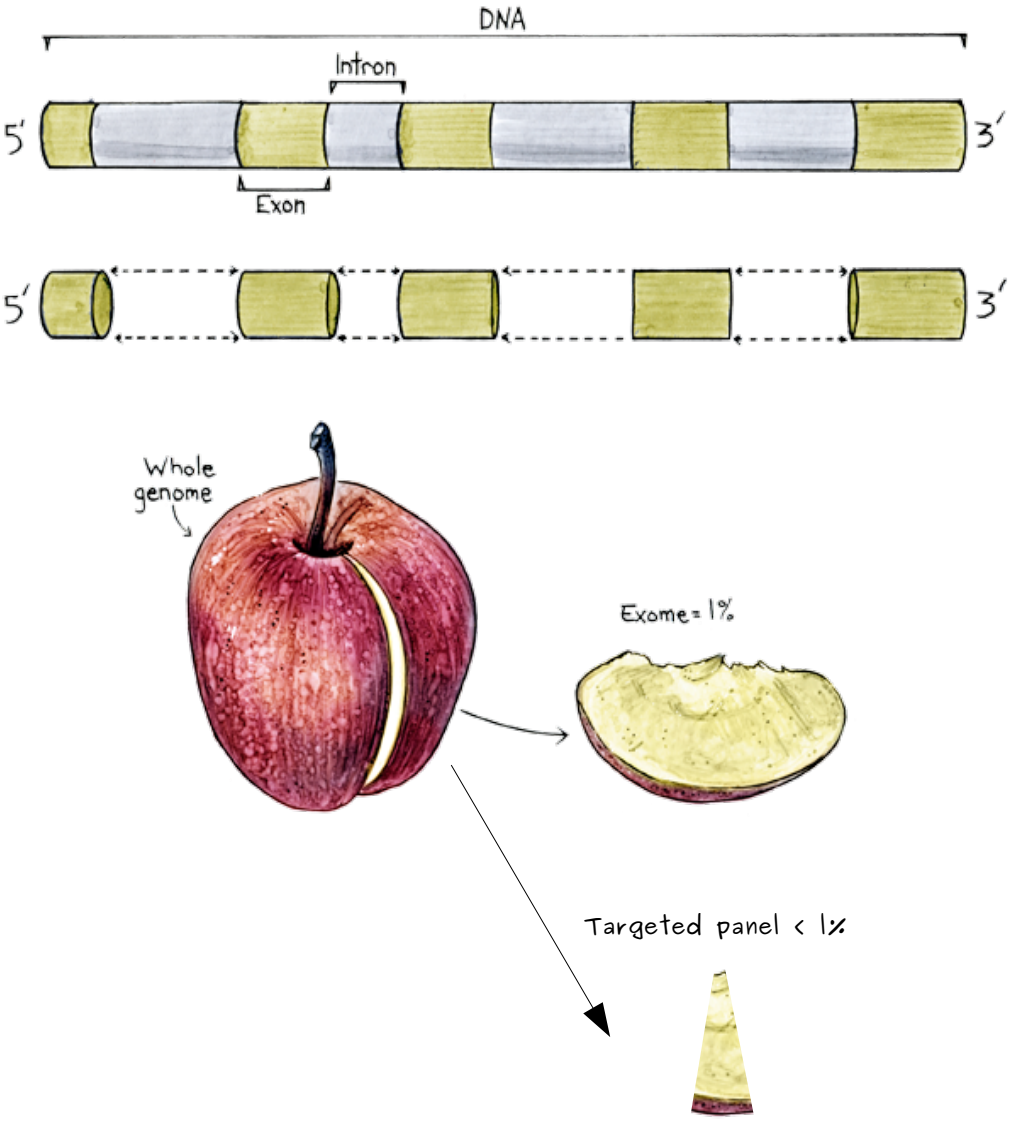
(…)

3. It is the responsibility of the ordering clinician/team to provide comprehensive pre- and posttest counseling to the patient.
   - Clinicians should be familiar with the basic attributes and limitations of clinical sequencing.
   - Clinicians should alert patients to the possibility that clinical sequencing may generate incidental findings that could require further evaluation.

(…)

5. The Working Group recommends that the ACMG, together with content experts and other professional organizations, refine and update this list at least annually.

Variant sequencing techniques



Adapted from: Copyright © 2012 University of Washington

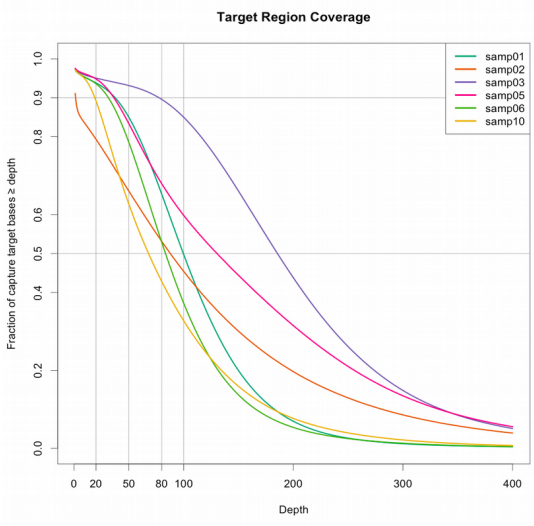
Depending on the aim (and funds!), different types of DNA sequencing can be chosen:

- **whole-genome**
- **exome**
- **targeted panel**

The amount of generated data can range from a few MB to several hundreds of GB!

Sequencing depth also plays an important role in identifying mutations.

For example, for cancer samples, it is recommended to use a coverage >= 100X.



Variant analysis workflow scheme – overview

1. Wet lab

- Genomic DNA extraction from sample
- Library preparation:
  - fragmentation
  - end-repair and A-tailing
  - adapter ligation
  - PCR amplification
- [for exome: target capture enrichment]
- Template preparation
- Sequencing

2. Bioinf. Analysis I

- Base calling
- De-multiplexing
- Quality control

3. Bioinf. Analysis II

- Adapter trimming, lane merging
- Read mapping to ref. genome
- Read deduplication

4. Bioinf. Analysis III

- Variant calling
- Variant annotation
- Variant filtering (prioritisation)
- Visual inspection (e.g. IGV)

5. Wet lab

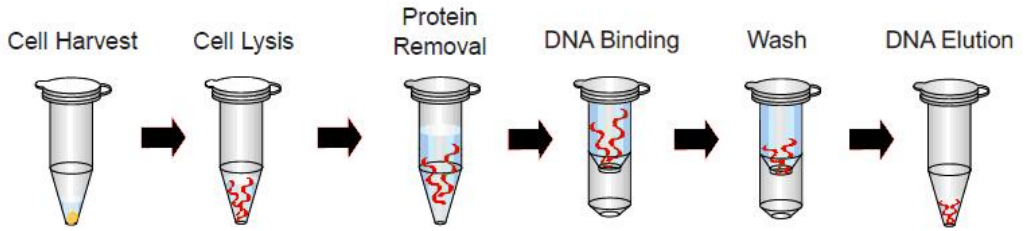
Confirmation by:

- Sanger sequencing
- MLPA (Multiplex Ligation-dependent Probe Amplification, a multiplex PCR method detecting copy numbers at 1bp resolution)

Variant analysis workflow scheme – DNA extraction

1. Wet lab

Genomic DNA extraction from sample



Variant analysis workflow scheme – library preparation

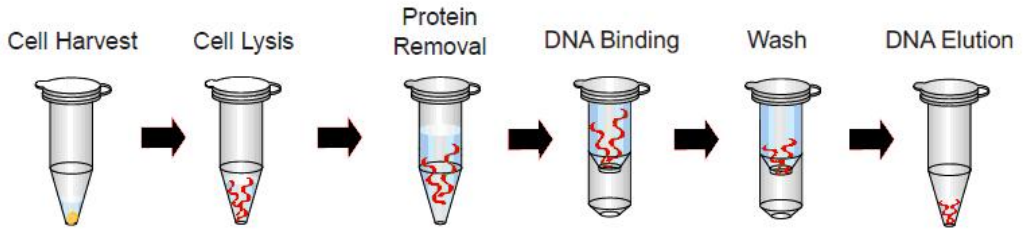
1. Wet lab

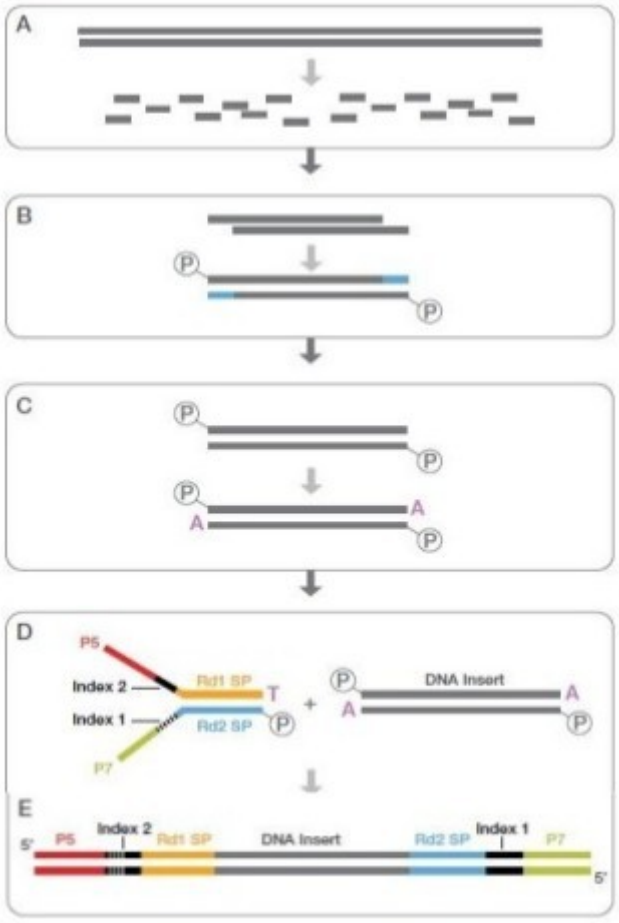
Genomic DNA extraction from sample

v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification





Variant analysis workflow scheme – target capture

1. Wet lab

Genomic DNA extraction from sample

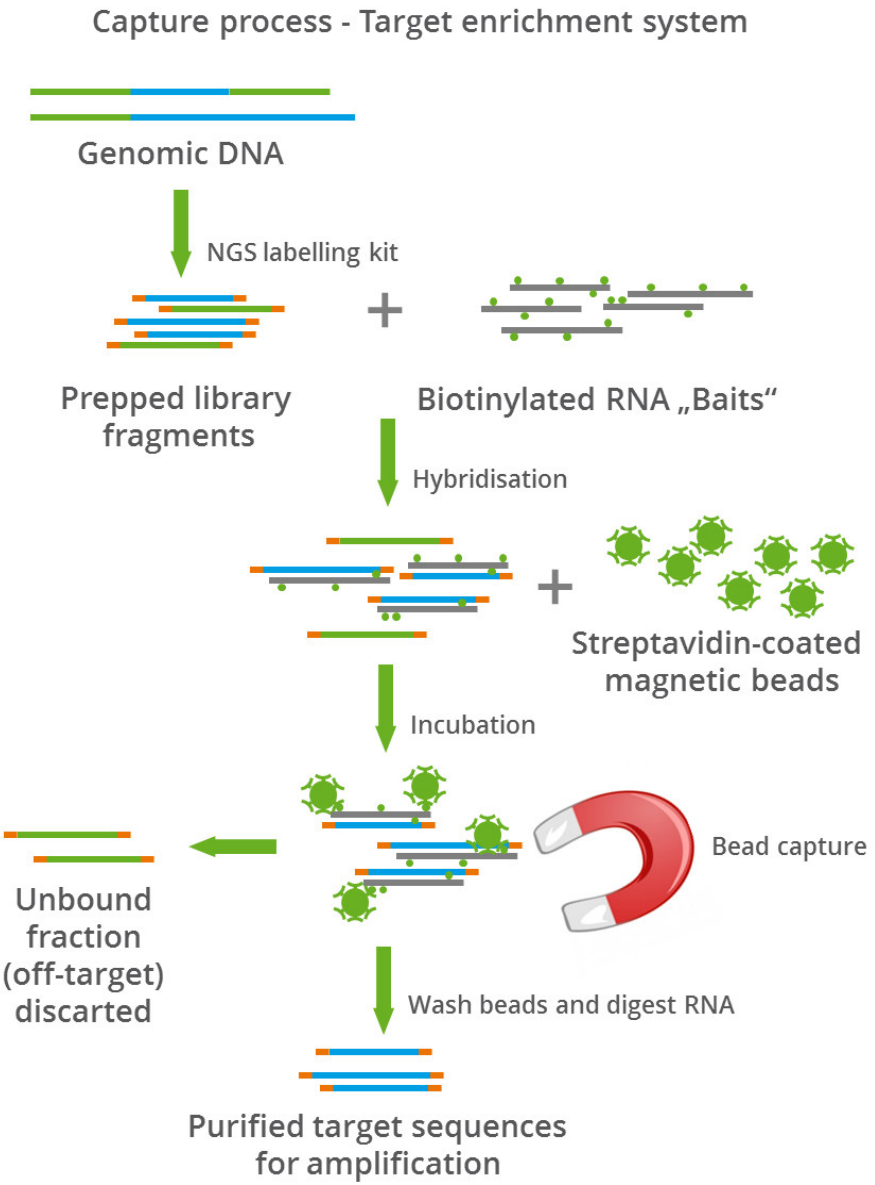
v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification

v

[for exome: target capture enrichment]



Variant analysis workflow scheme – template preparation

1. Wet lab

Genomic DNA extraction from sample

v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification

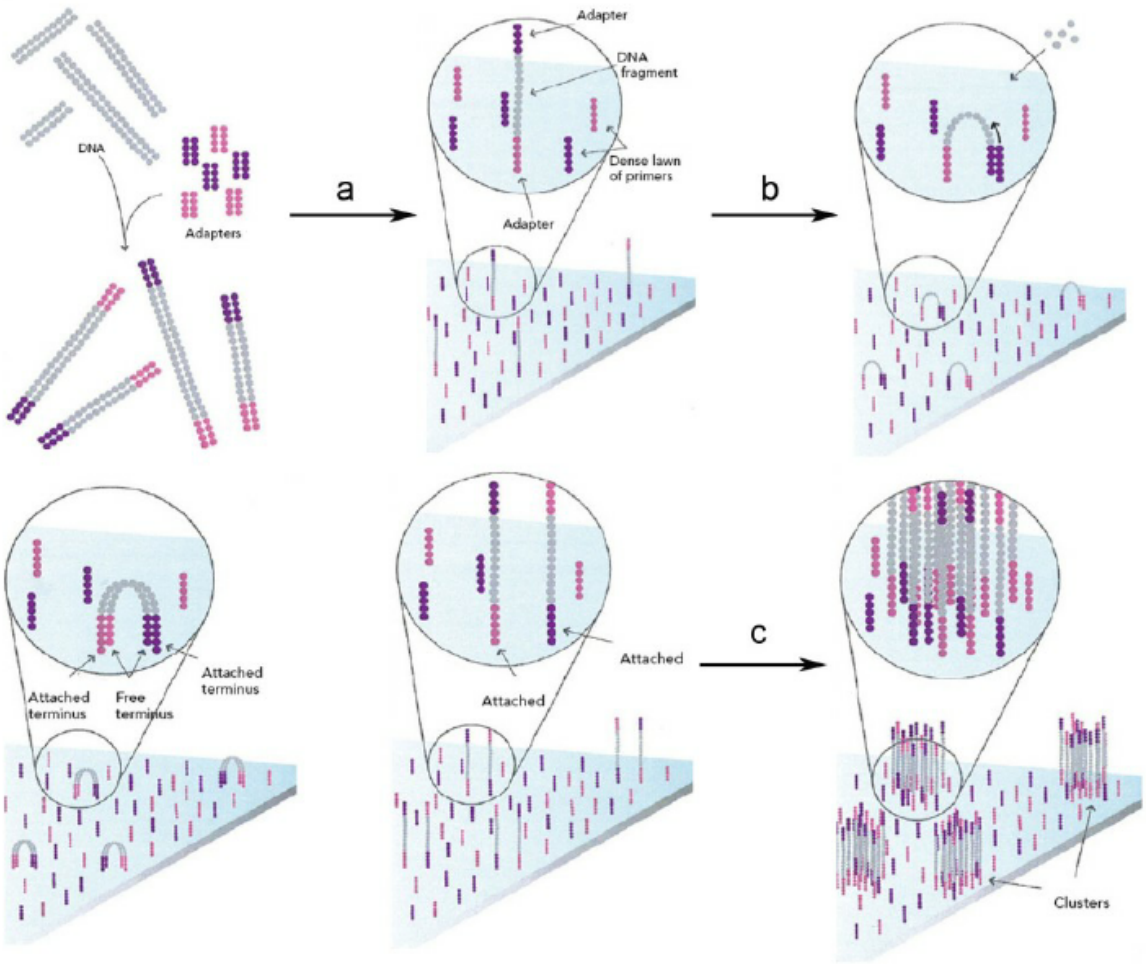
v

[for exome: target capture enrichment]

v

Template preparation

Illumina's strategy: read clusters formation by bridge amplification



Variant analysis workflow scheme – sequencing strategies

1. Wet lab

Genomic DNA extraction from sample

v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification

v

[for exome: target capture enrichment]

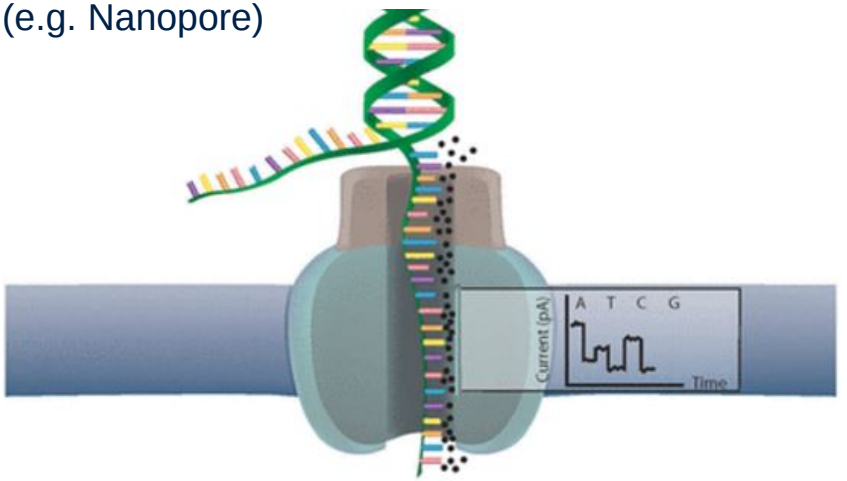
v

Template preparation

v

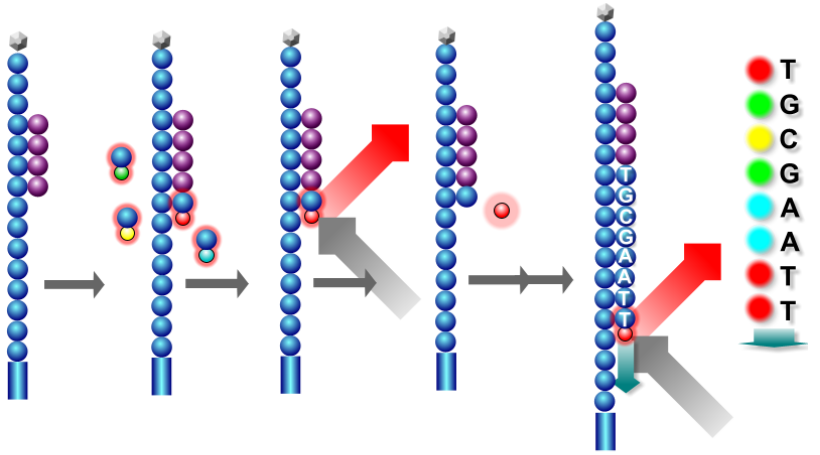
Sequencing

Seq-by-Scan (e.g. Nanopore)



Seq-by-Syn (e.g. Illumina, PacBio, IonTorrent)

Detecting: fluorescence pH



Variant analysis workflow scheme – base calling and BCL file format

1. Wet lab

Genomic DNA extraction from sample

v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification

v

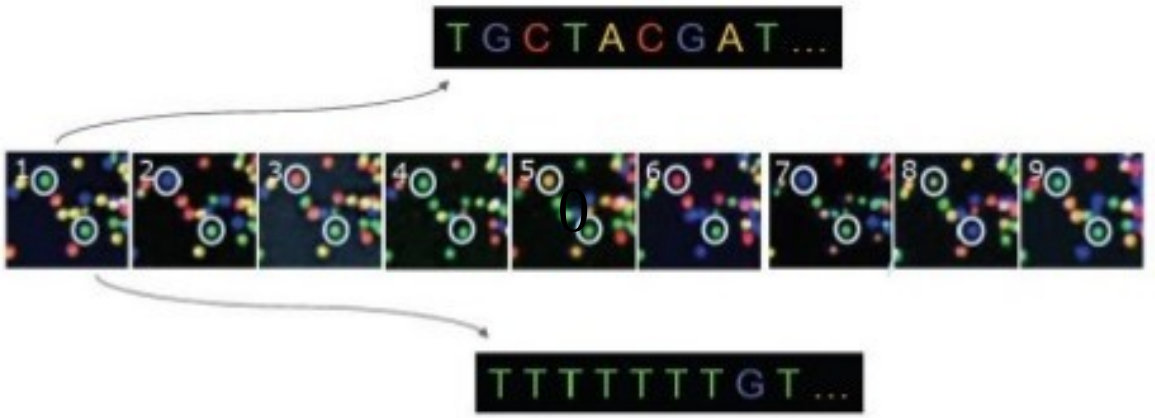
[for exome: target capture enrichment]

v

Template preparation

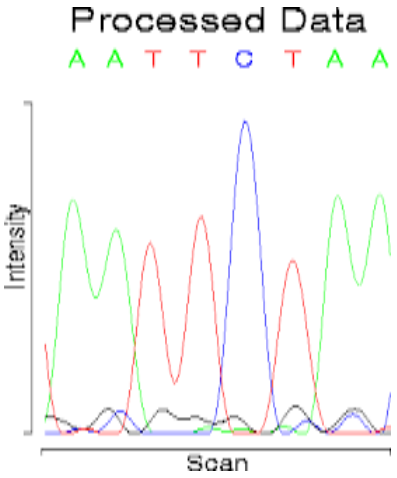
v

Sequencing



2. Bioinf. Analysis I

Base calling



This information gets written “on the fly” (as opposed to waiting until the run is over and calling bases for the entire read) into **BCL files** (“.bcl”), binary files containing **base calls and qualities** for each flowcell's tile in each cycle.

Variant analysis workflow scheme – de-multiplexing

1. Wet lab

Genomic DNA extraction from sample

v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification

v

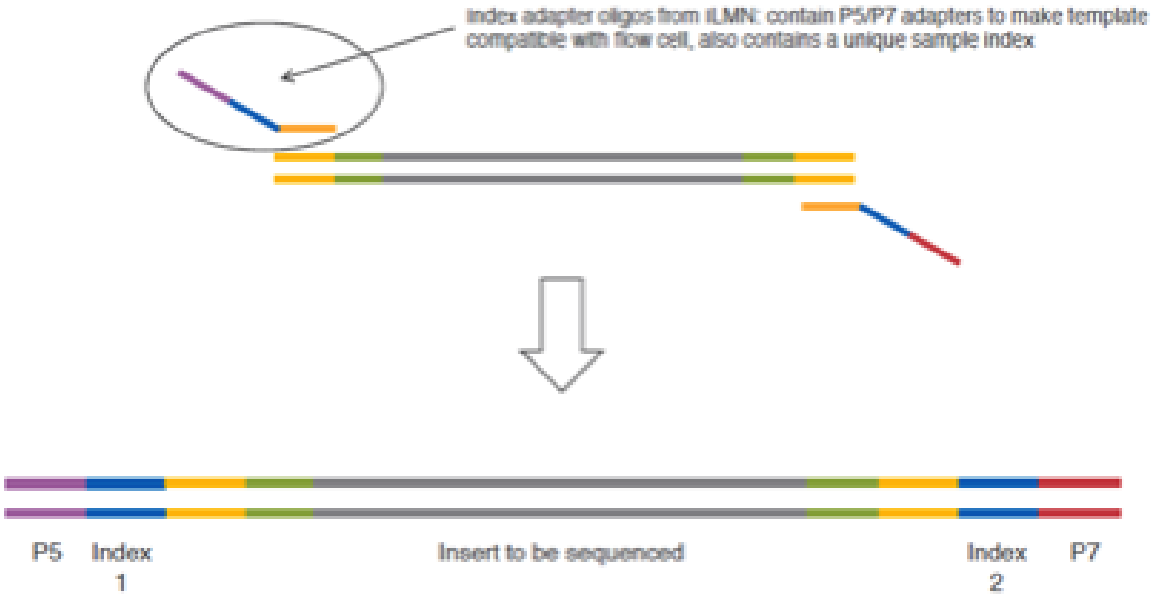
[for exome: target capture enrichment]

v

Template preparation

v

Sequencing

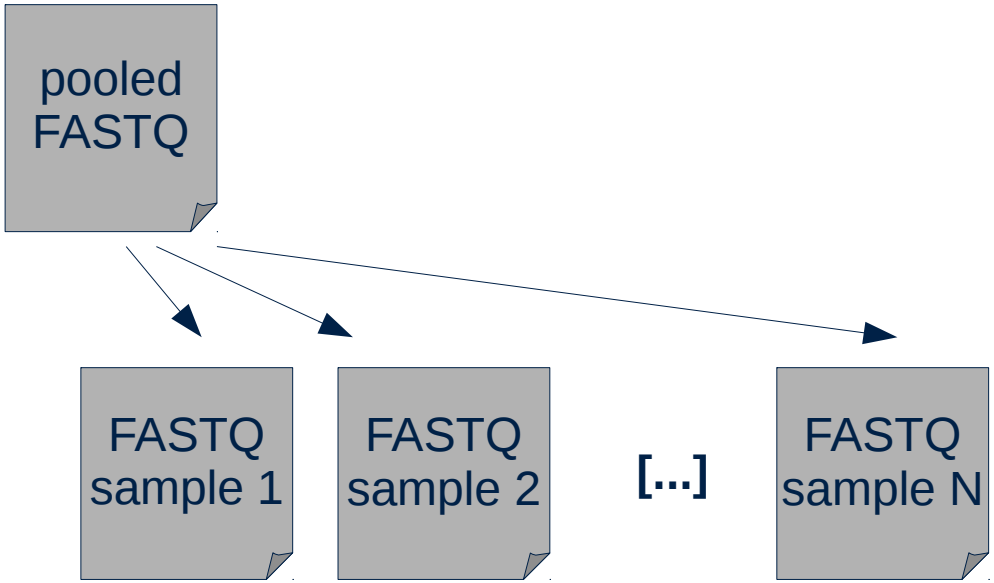


2. Bioinf. Analysis I

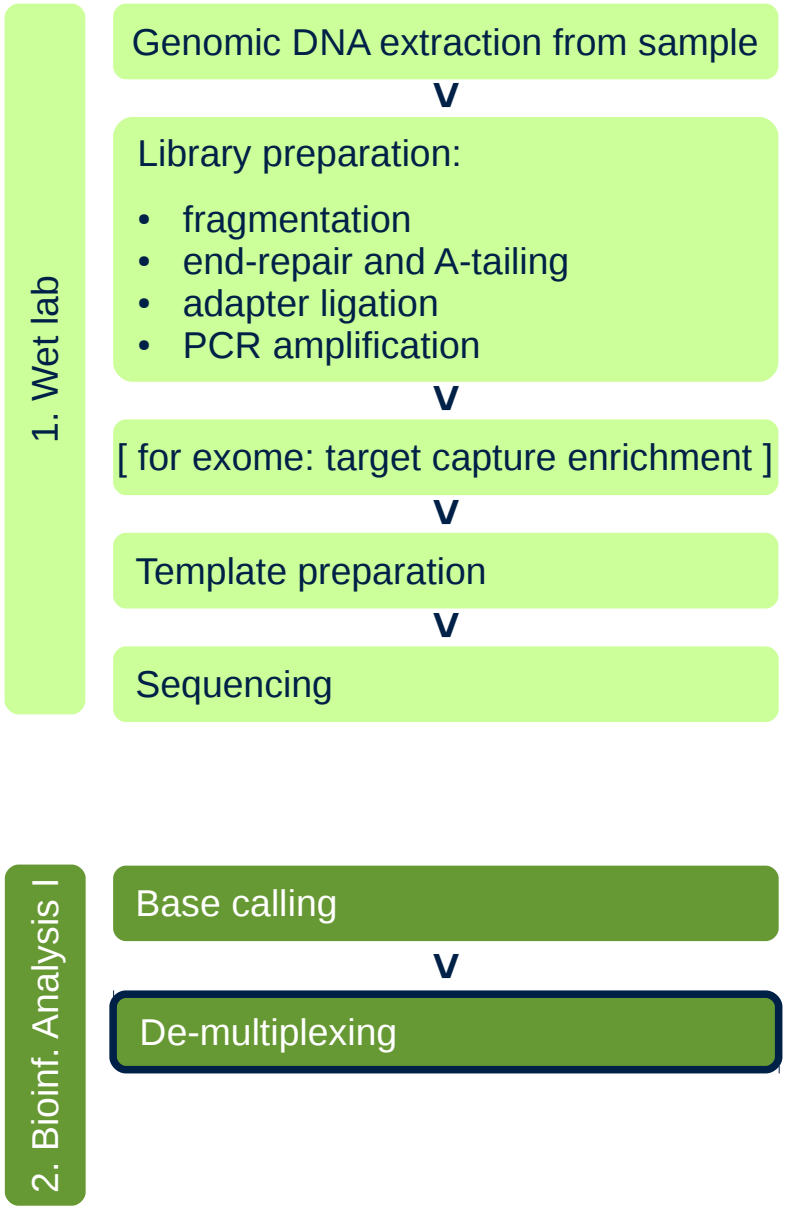
Base calling

v

De-multiplexing



Variant analysis workflow scheme – FASTQ file format



The **FASTQ** format is a text-based file format to store both biological sequences (DNA or RNA) and their corresponding quality scores. It is usually generated from a BCL file. Both the sequence letter and the quality score are encoded with a single ASCII character for brevity. A FASTQ file normally uses four lines per sequence:

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description
- Line 2 is the raw sequence letters
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description)
- Line 4 encodes the quality values for the sequence in Line 2 (therefore must be as long as Line 2).

Example of a FASTQ file containing a single sequence read:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Here are the quality value characters in left-to-right increasing order of quality (ASCII):

```
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ
[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

E.g. Python code to convert ASCII to number and viceversa:

```
>>> ord("%")
37
>>> chr(37)
'%'
```

Variant analysis workflow scheme – quality control

1. Wet lab

Genomic DNA extraction from sample

v

Library preparation:

- fragmentation
- end-repair and A-tailing
- adapter ligation
- PCR amplification

v

[for exome: target capture enrichment]

v

Template preparation

v

Sequencing

2. Bioinf. Analysis I

Base calling

v

De-multiplexing

v

Quality control

It is useful to have a look at the overall quality of the reads before and after trimming, e.g. using **FastQC** :