



Variant analysis course

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

13.11.2019

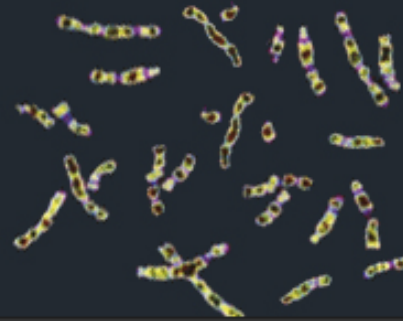
Table of contents

- Overview of the major genomic projects
- Current sequencing techniques and features
- Mutation types and downstream effects
- Incidental findings
- From DNA extraction to library sequencing
- Base calling, de-multiplexing and QC
- Read filtering, mapping and de-duplication
- Variant calling and annotation
- Variant filtering and visualisation
- Variant validation in the lab
- Extra steps: computational phenotype analysis, linkage and GWAS

Major genomic projects – 1000 Genomes Project

1000 Genomes

A Deep Catalog of Human Genetic Variation



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.

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

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:

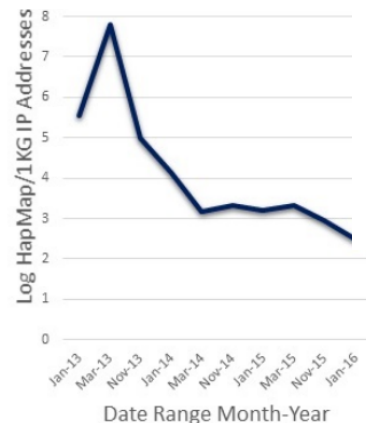
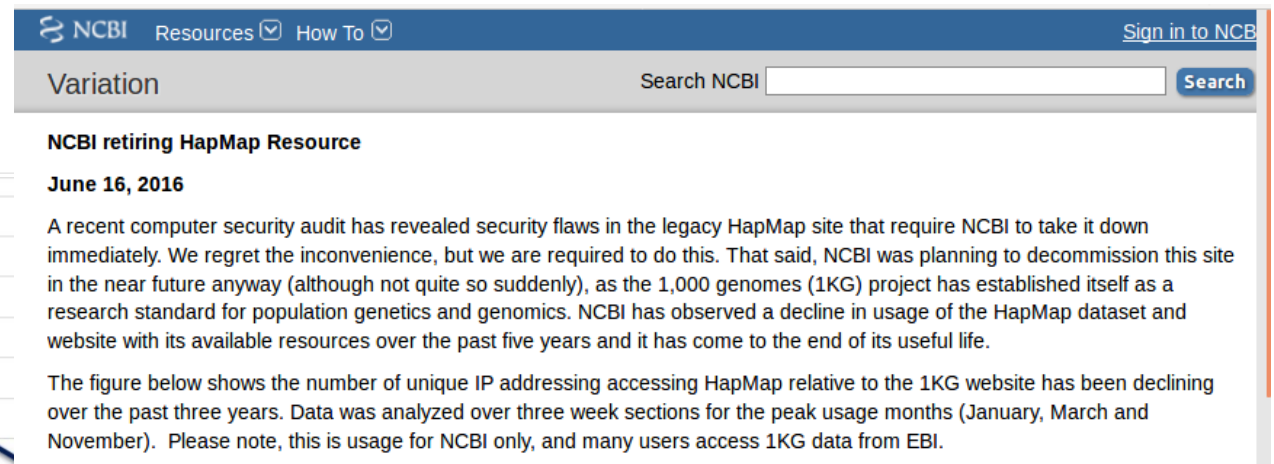


Figure 1: HapMap vs 1KG usage over the past three years exemplified during the peak usage months.

Major genomic projects – UK10K



UK10K

Rare Genetic Variants in Health and Disease

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



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.

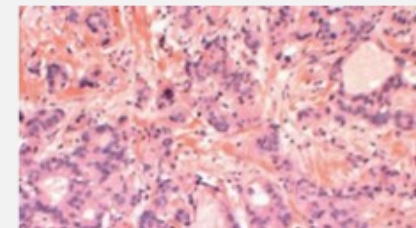
The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

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.



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 seminars

[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



Major genomic projects – ExAC

ExAC Browser Beta

Gene, transcript, variant, or i

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 60,766 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 Ardlissino
- 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 Klezun
- 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: MacArthur and Neale)

Interested in working on the development of this resource? [Apply here](#).

gnomAD browser beta | genome Aggregation Database

Search for a gene or variant or region

Example - Gene: [PCSK9](#), Variant: [1-55516888-G-GA](#)

About gnomAD

The [Genome Aggregation Database](#) (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans [123,136 exome sequences and 15,496 whole-genome sequences](#) from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed [here](#).

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use [here](#).

Sign up for our mailing list for future release announcements [here](#).

Recent News

October 3, 2017

[gnomAD r2.0.2](#) released. Sample composition is identical to the previous release (r2.0.1), however we have made a change to the variant filtering process that you can read about [here](#).

February 27, 2017

[Official gnomAD release \(version 2.0\)](#) with browser updates and data available for [download](#).

October 19, 2016

Public release of gnomAD Browser (beta) at ASHG!

Major genomic projects – POBI



People of the British Isles

[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](#)

Country	Visits
United Kingdom	97

1-1 of 1

7



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



Top 10 Regions	Count
Cumbria	316
Corwall	309
Yorkshire	289
North East	274
Orkney	255
Sussex/Kent	190
Oxfordshire	187
Lincolnshire	181
Northern Ireland	175
Norfolk	100

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  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)



Search

Home About Cancer Genomics Cancers Selected for Study Research Highlights Publications

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 publicly 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 publicly available.

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over

2.5
PETABYTES
of data

To put this into perspective, 1 petabyte of data is equal to

212,000
DVDs

TCGA data describes

33
DIFFERENT
TUMOR TYPES

including

10
RARE
CANCERS

based on paired tumor and normal tissue sets

11,000
PATIENTS

using

7
DIFFERENT
DATA TYPES

80

TCGA RESULTS & FINDINGS



Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the breast-like subtype of breast cancer to be similar to the serious subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.



Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations*



Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM

20
COLLABORATING
INSTITUTIONS
across the United States and Canada

WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

*TCGA studies of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by mutations in Epstein-Barr virus.

www.cancer.gov/ccg

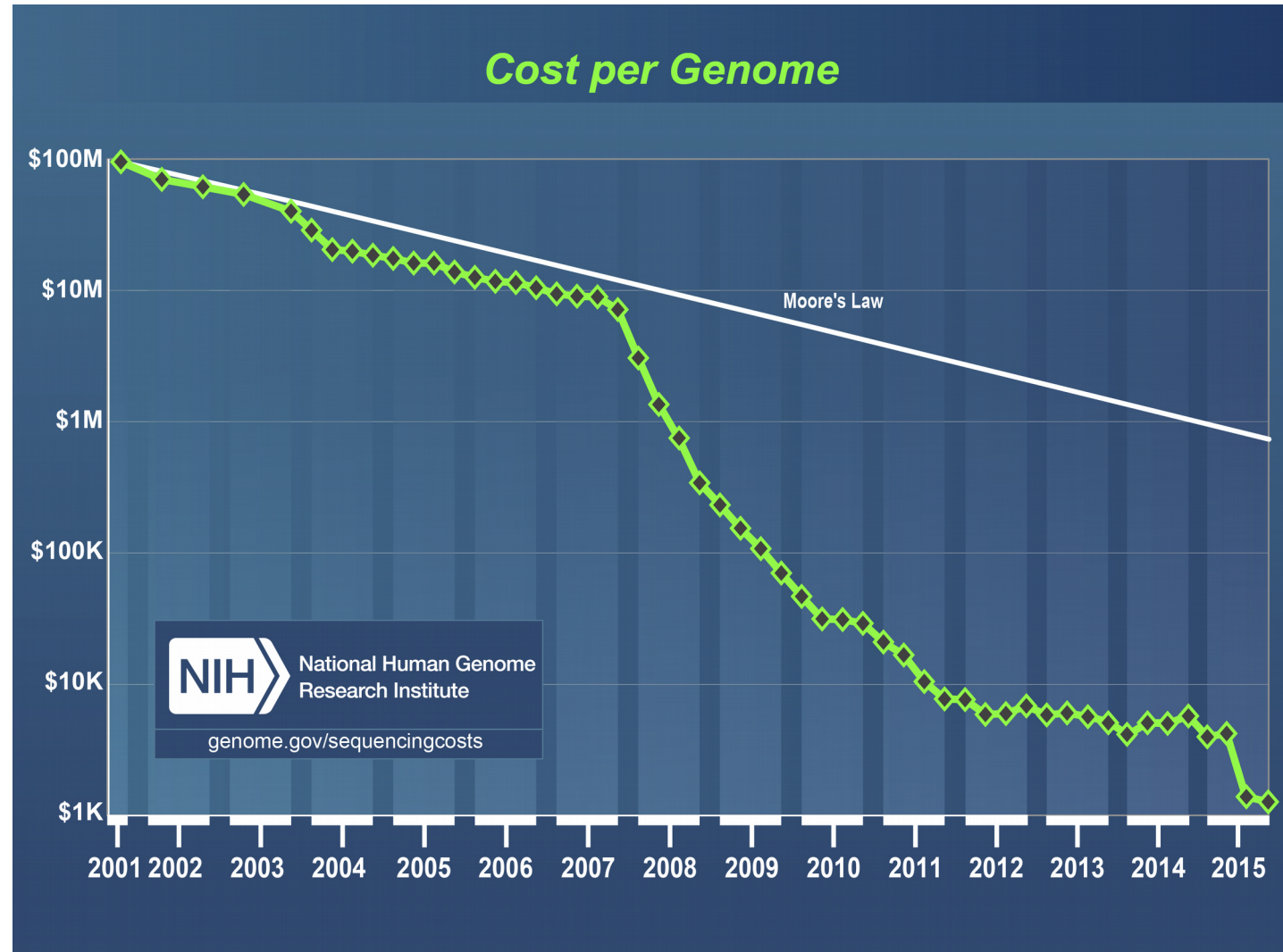
TCGA is the biggest component of the International Cancer Genome Consortium (ICGC), a collaboration of scientists from 16 nations that has discovered nearly 10 million cancer-related mutations.

Recent advances in sequencing technology

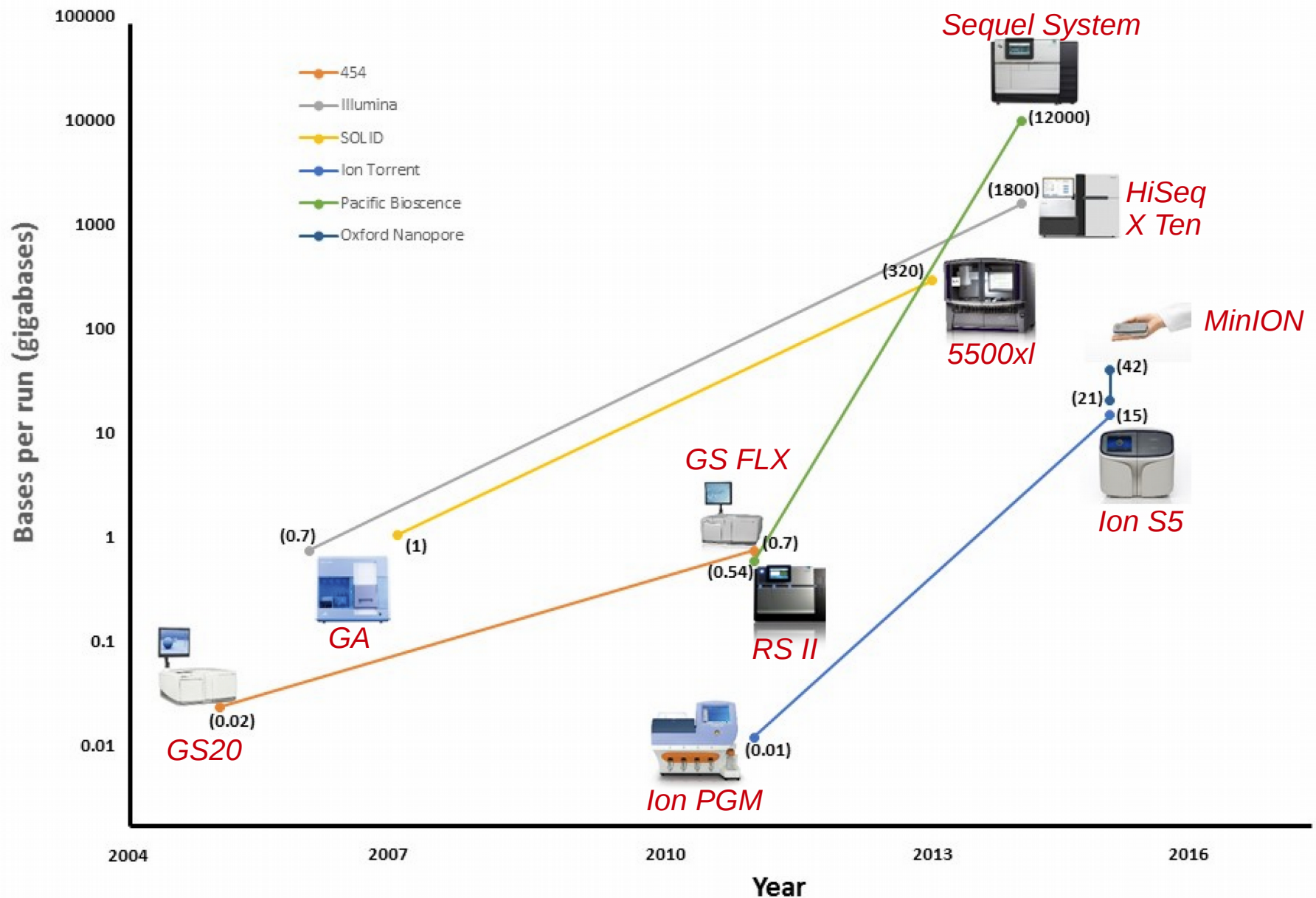
The sequencing cost has significantly dropped during the last decade, with a 10'000-fold reduction that made it possible to reach the target of sequencing a human genome for USD \$ 1'000 (announced by Illumina in 2014, through the HiSeq X Ten machine), thus marking the beginning of a new era of personalized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



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

Reduction of

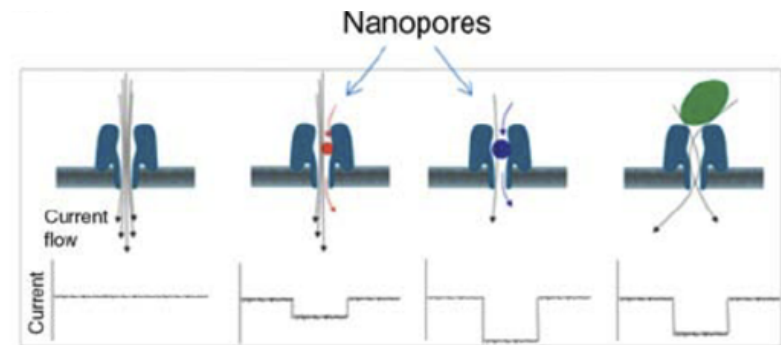
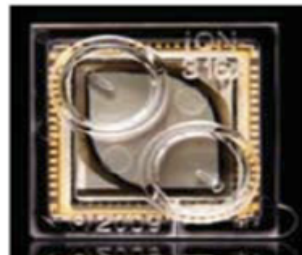
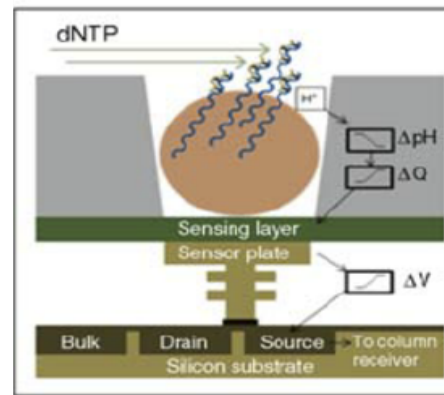
- Sequencing cost
- DNA input material
- Workflow duration
- Platform size

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)

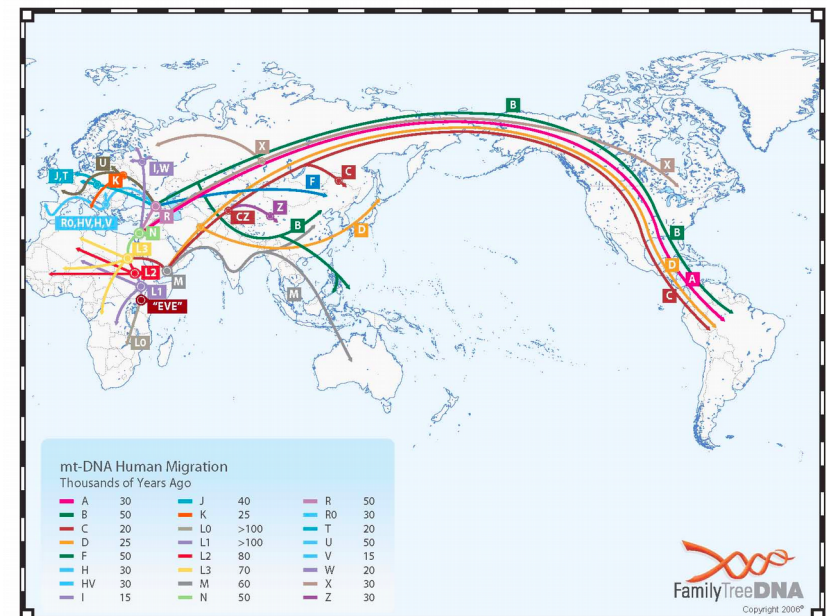
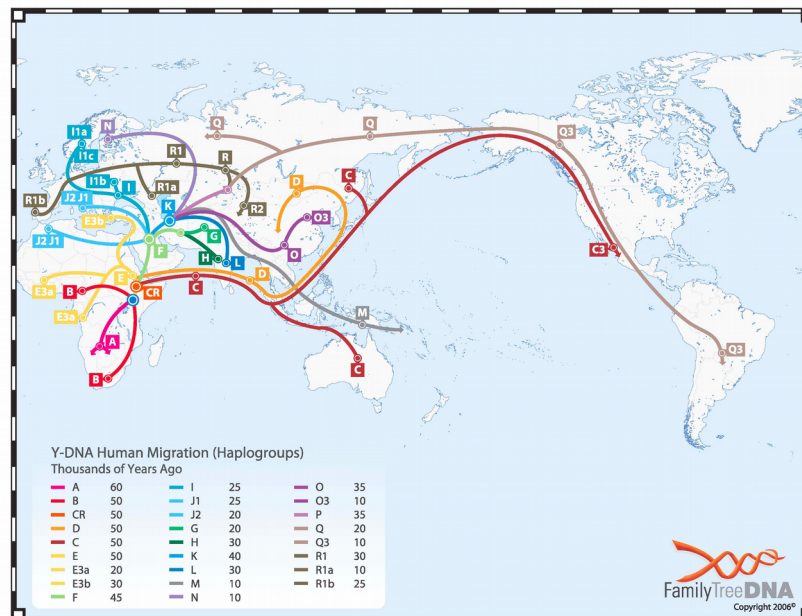
Genetic variation – genotype and haplotype

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).

		pollen ♂	
		B	b
pistil ♀	B	BB	Bb
	b	Bb	bb

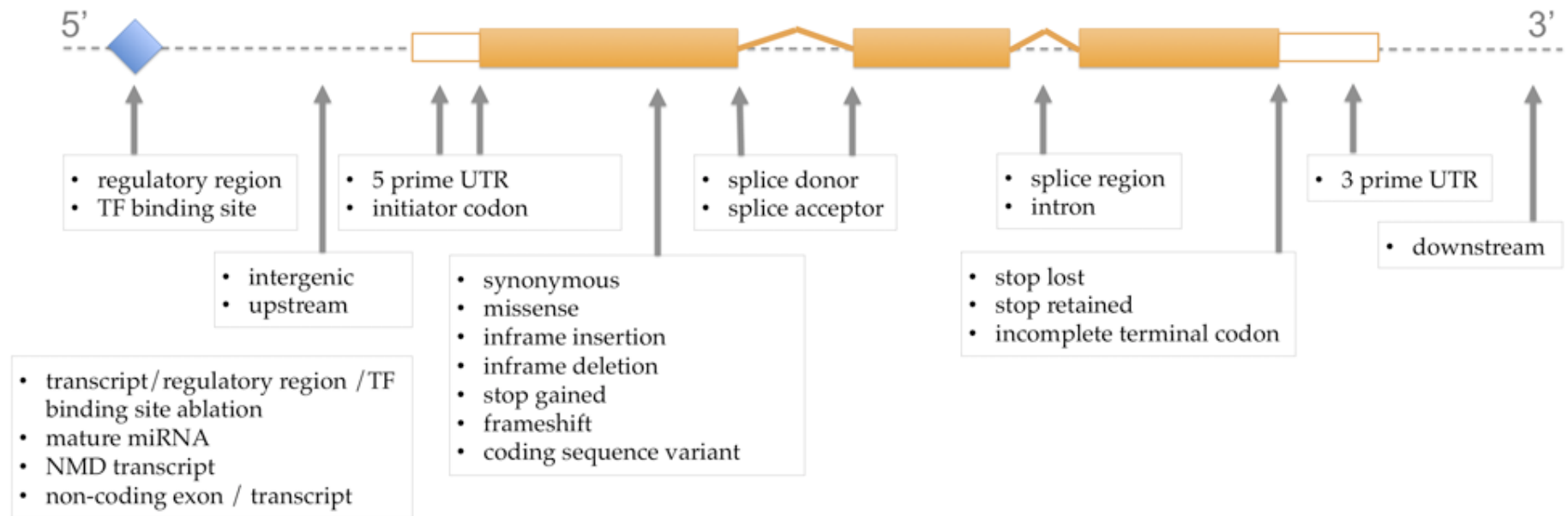
Haplotype: set of DNA variations (SNPs or STRs) that are usually inherited together from a single parent, being located on the same chromosome. The alleles making up a haplotype can be located in different places on the chromosome.

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.

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

Robert C. Green, MD, MPH^{1,2}, Jonathan S. Berg, MD, PhD³, Wayne W. Grody, MD, PhD⁴⁻⁶, Sarah S. Kalia, ScM, CGC¹, Bruce R. Korf, MD, PhD⁷, Christa L. Martin, PhD, FACMG⁸, Amy L. McGuire, JD, PhD⁹, Robert L. Nussbaum, MD¹⁰, Julianne M. O'Daniel, MS, CGC³, Kelly E. Ormond, MS, CGC¹¹, Heidi L. Rehm, PhD, FACMG^{2,12}, Michael S. Watson, PhD, FACMG¹³, Marc S. Williams, MD, FACMG¹⁴ and Leslie G. Biesecker, MD¹⁵

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	<i>BRCA1</i>	113705	AD	KP and EP
				<i>BRCA2</i>	600185		
Li-Fraumeni syndrome	151623	20301488	Child/adult	<i>TP53</i>	191170	AD	KP and EP
Peutz-Jeghers syndrome	175200	20301443	Child/adult	<i>STK11</i>	602216	AD	KP and EP
Lynch syndrome	120435	20301390	Adult	<i>MLH1</i>	120436	AD	KP and EP
				<i>MSH2</i>	609309		
				<i>MSH6</i>	600678		
				<i>PMS2</i>	600259		
Familial adenomatous polyposis	175100	20301519	Child/adult	<i>APC</i>	611731	AD	KP and EP
<i>MYH</i> -associated polyposis; adenomas, multiple colorectal, <i>FAP</i> type 2; colorectal adenomatous polyposis, autosomal recessive, with pilomatricomas	608456 132600	23035301	Adult	<i>MUTYH</i>	604933	AR ^c	KP and EP
Von Hippel-Lindau syndrome	193300	20301636	Child/adult	<i>VHL</i>	608537	AD	KP and EP
Multiple endocrine neoplasia type 1	131100	20301710	Child/adult	<i>MEN1</i>	613733	AD	KP and EP
Multiple endocrine neoplasia type 2	171400 162300	20301434	Child/adult	<i>RET</i>	164761	AD	KP

(...) (...) (...) (...) (...) (...) (...)

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.

(...)

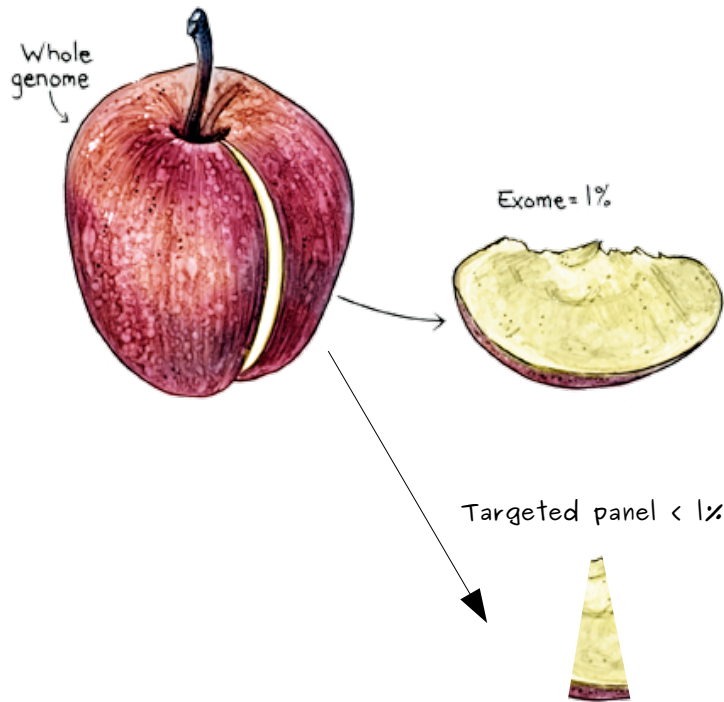
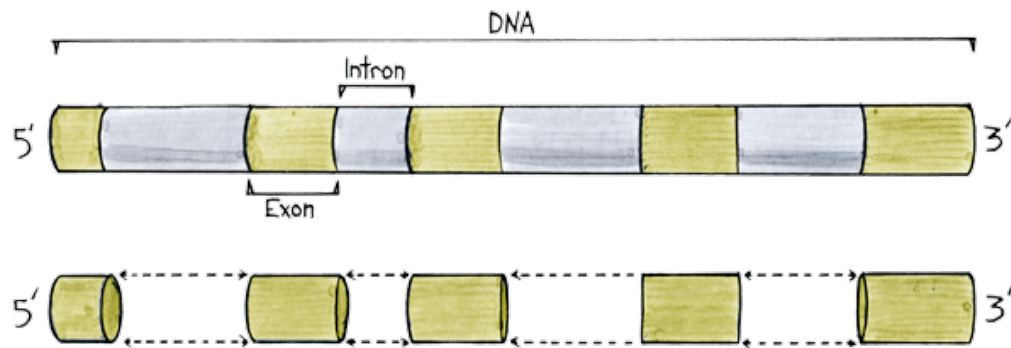
3. It is the responsibility of the ordering clinician/team to provide comprehensive pre- and posttest counseling to the patient.
 - Clinicians should be familiar with the basic attributes and limitations of clinical sequencing.
 - Clinicians should alert patients to the possibility that clinical sequencing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

NOTE: KP means known pathogenic, EP means expected pathogenic

Variant sequencing techniques



Adapted from: Copyright © 2012 University of Washington

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

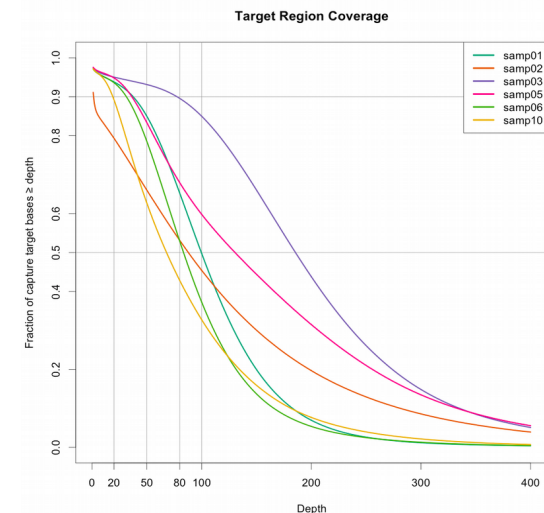
- whole-genome
- exome
- targeted panel

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

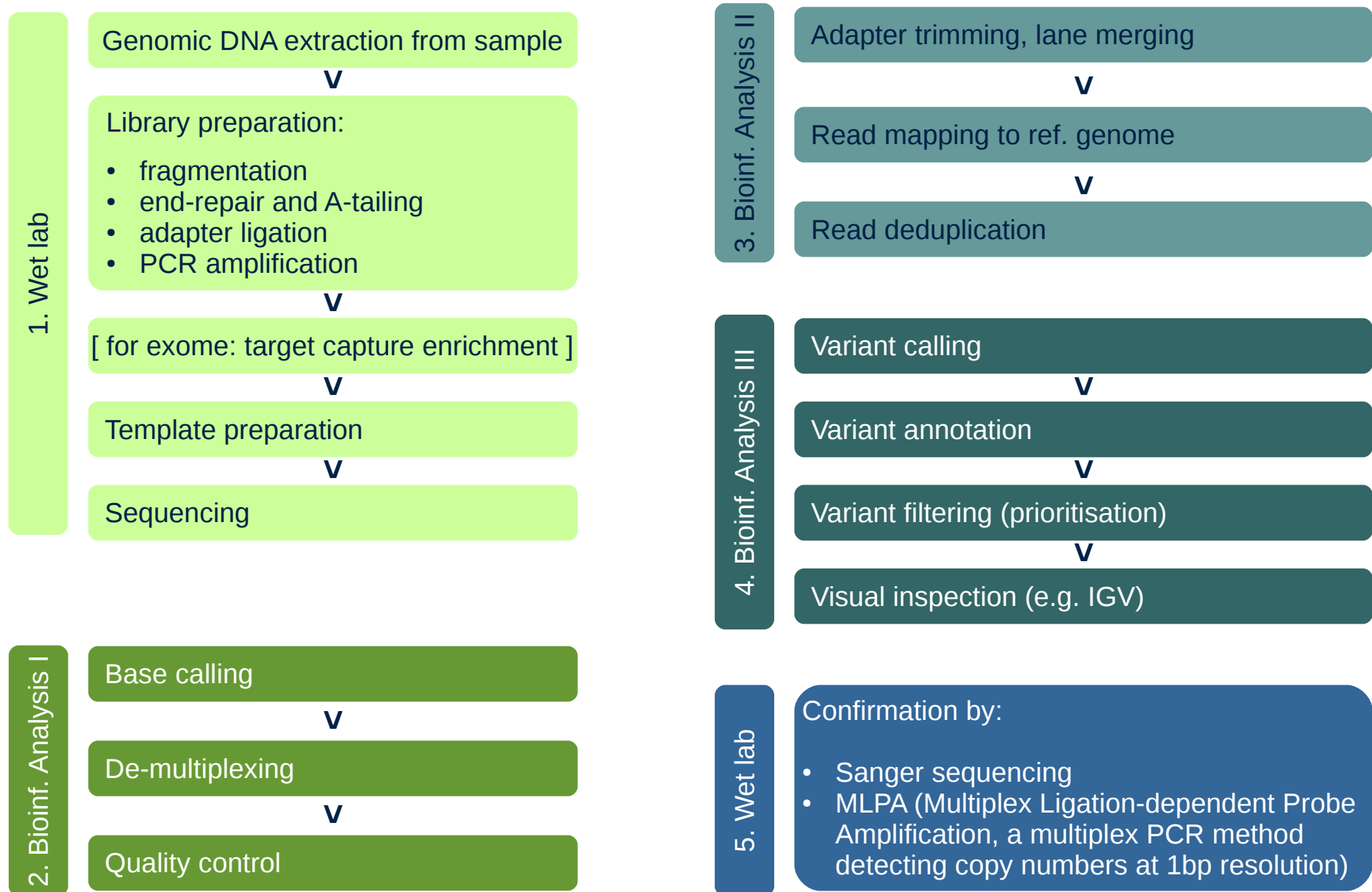


Sequencing depth also plays an important role in identifying mutations.

For example, for cancer samples, it is recommended to use a coverage $\geq 100\times$.



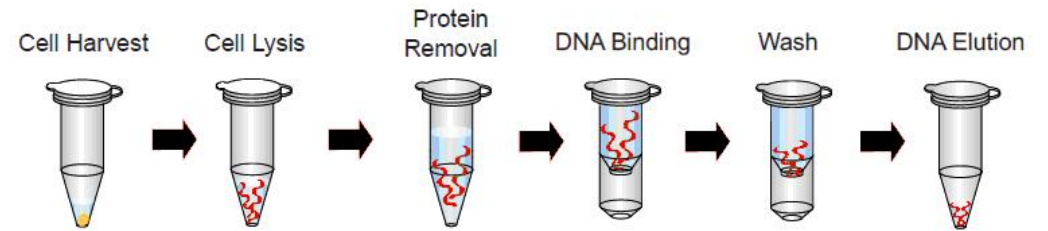
Variant analysis workflow scheme – overview



Variant analysis workflow scheme – DNA extraction

1. Wet lab

Genomic DNA extraction from sample



Variant analysis workflow scheme – library preparation

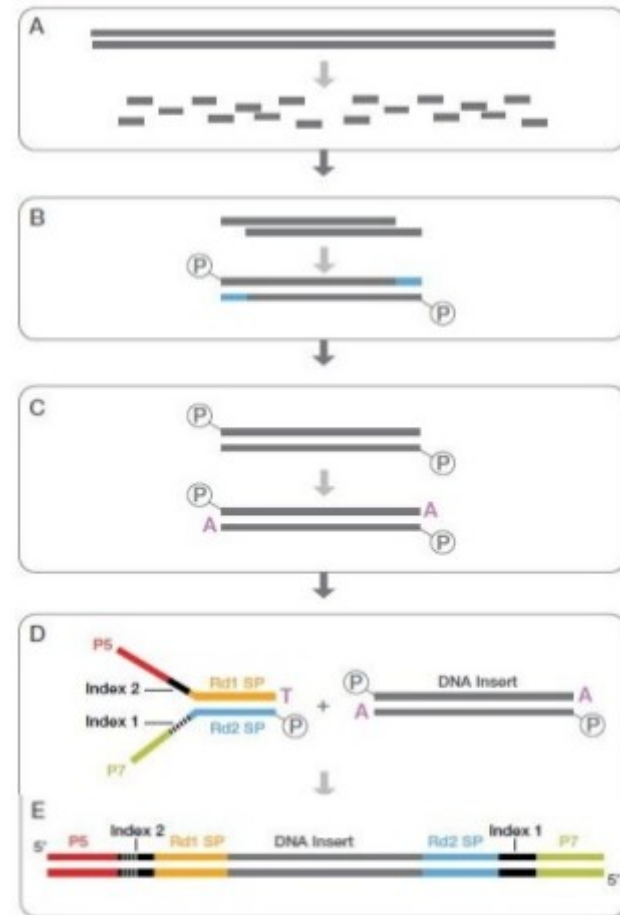
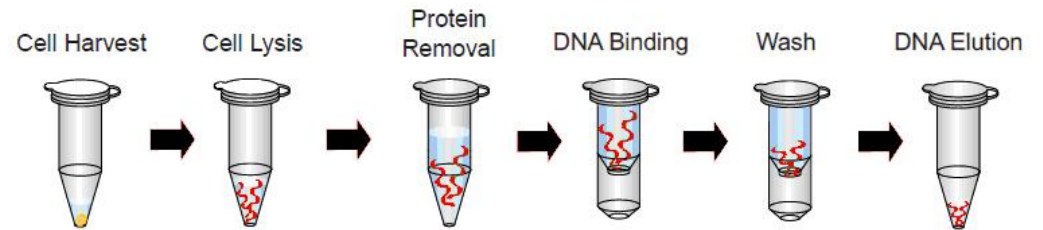
1. Wet lab

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



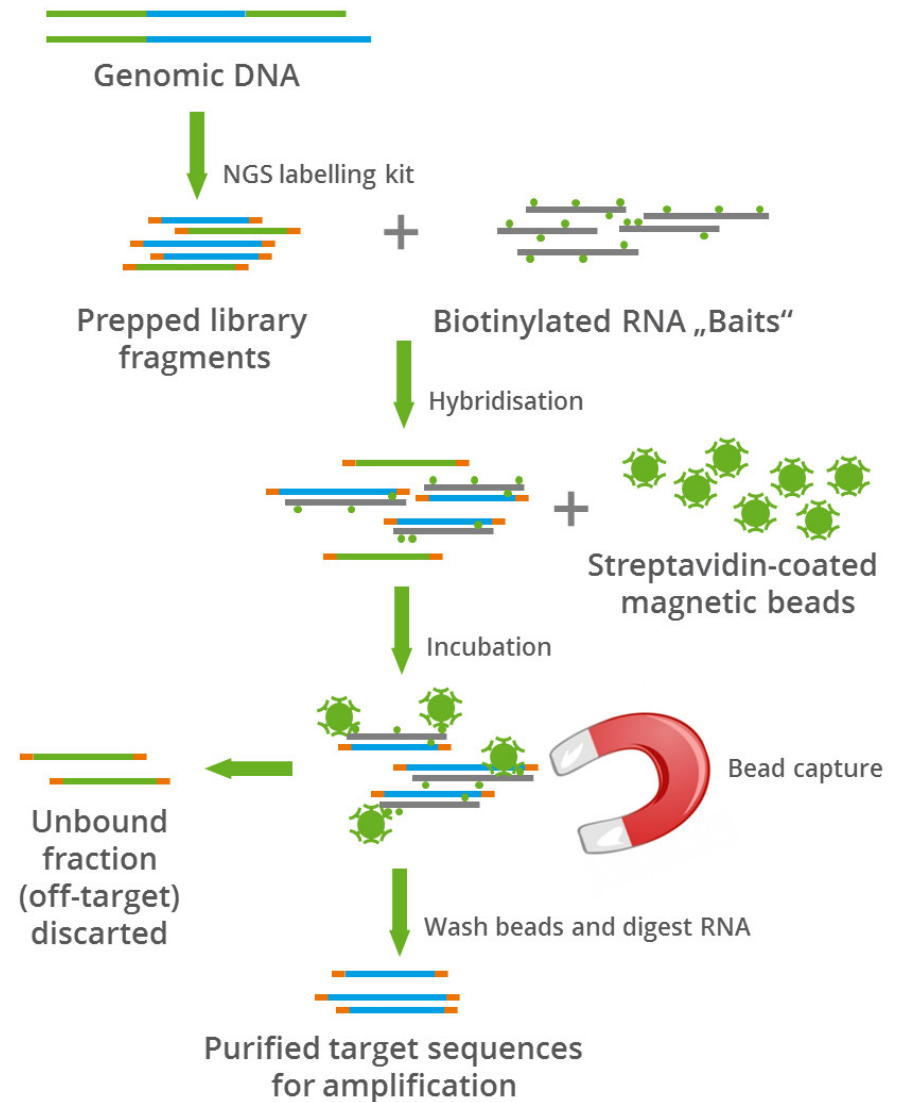
Library preparation:

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



[for exome: target capture enrichment]

Capture process - Target enrichment system



Variant analysis workflow scheme – template preparation

1. Wet lab

Genomic DNA extraction from sample

▼

Library preparation:

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

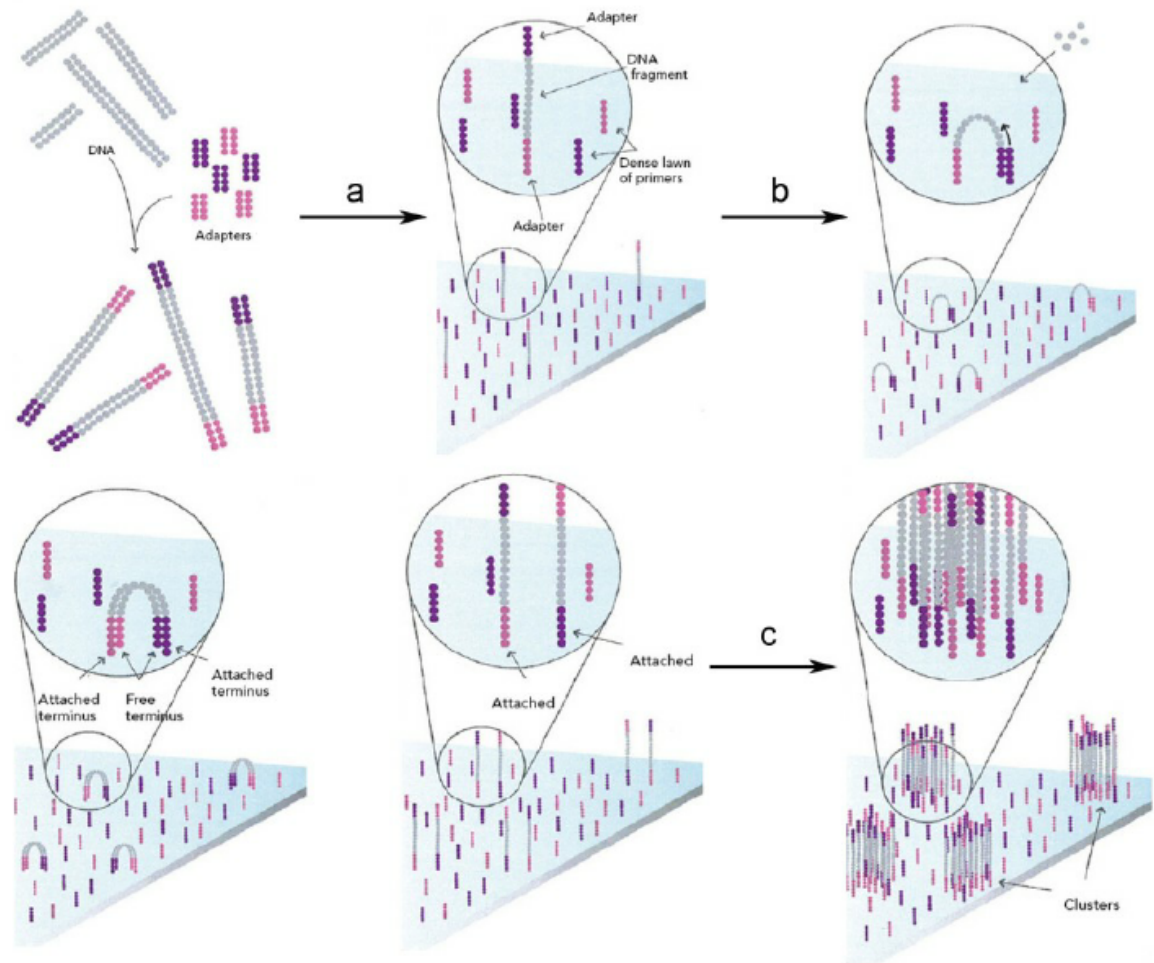
▼

[for exome: target capture enrichment]

▼

Template preparation

Illumina's strategy: read clusters formation by bridge amplification



Variant analysis workflow scheme – sequencing strategies

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]

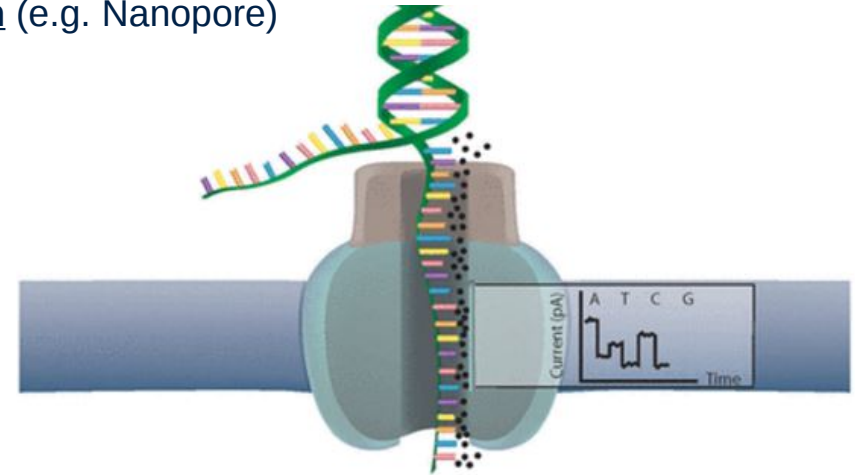


Template preparation



Sequencing

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

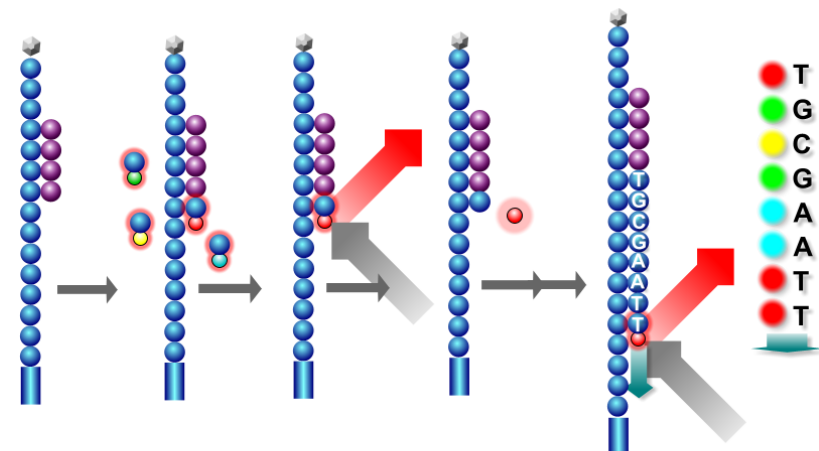


Seq-by-Syn (e.g. Illumina, PacBio, IonTorrent)

Detecting:

fluorescence

pH



Variant analysis workflow scheme – base calling and BCL file format

1. Wet lab

Genomic DNA extraction from sample



Library preparation:

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



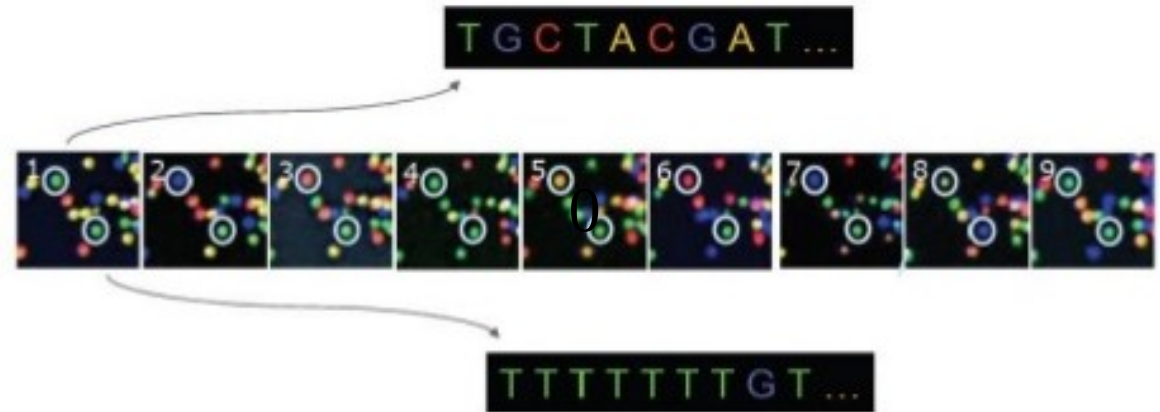
[for exome: target capture enrichment]



Template preparation

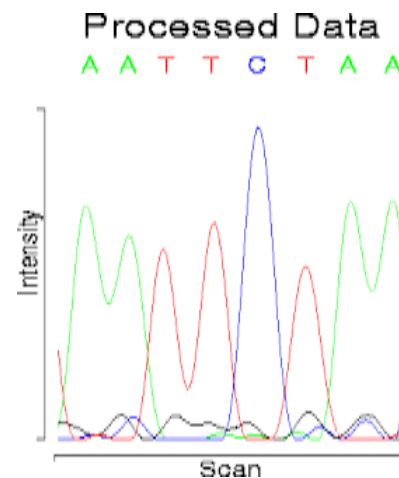


Sequencing



2. Bioinf. Analysis I

Base calling



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



Library preparation:

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



[for exome: target capture enrichment]



Template preparation



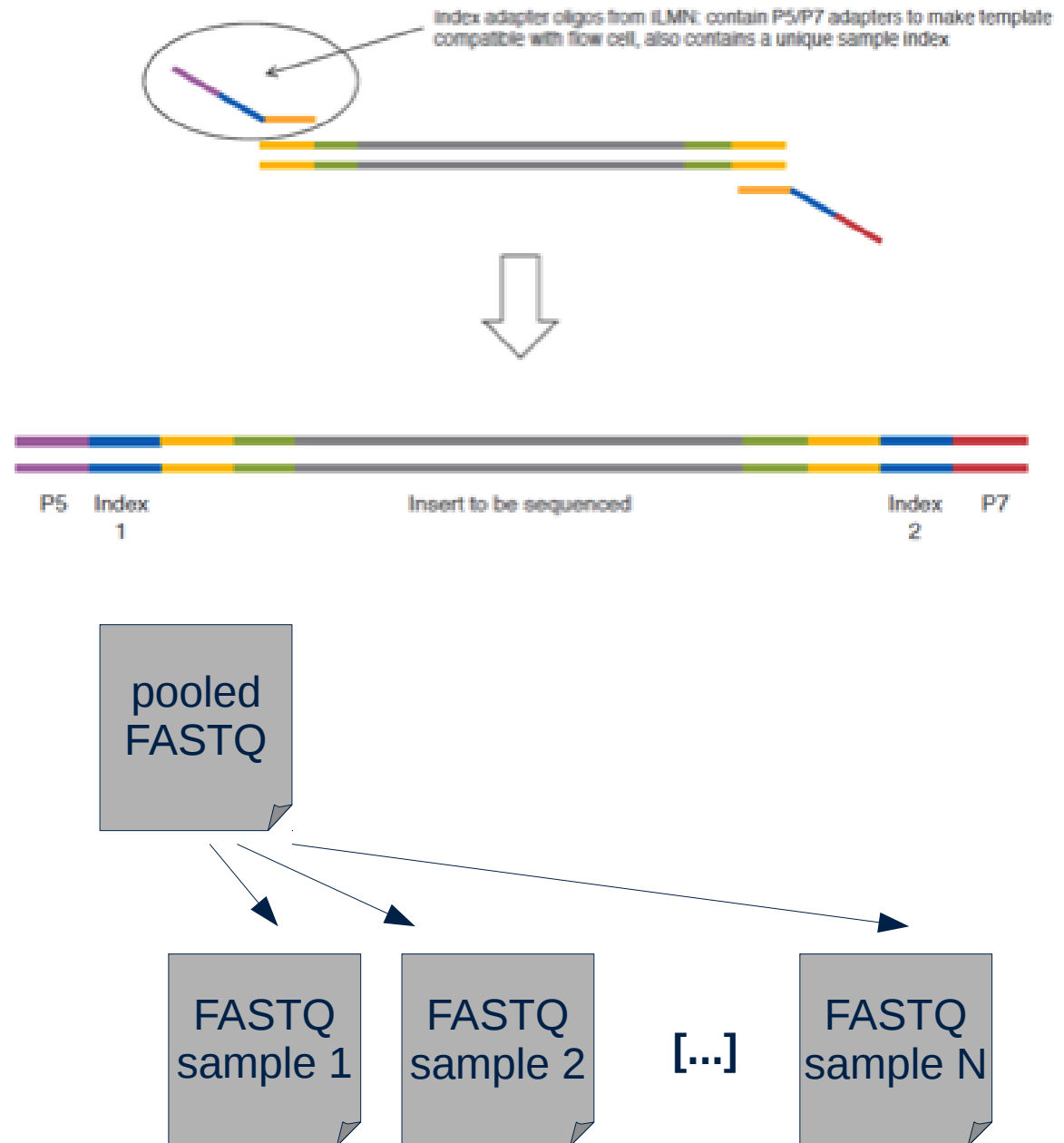
Sequencing

2. Bioinf. Analysis I

Base calling



De-multiplexing



Variant analysis workflow scheme – FASTQ file format

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

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
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCACACTCACAGTTT
+
!''*((( (**+))%%++) (%%%) .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)
'%'
```

2. Bioinf. Analysis I

Base calling



De-multiplexing

Variant analysis workflow scheme – quality control

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

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

