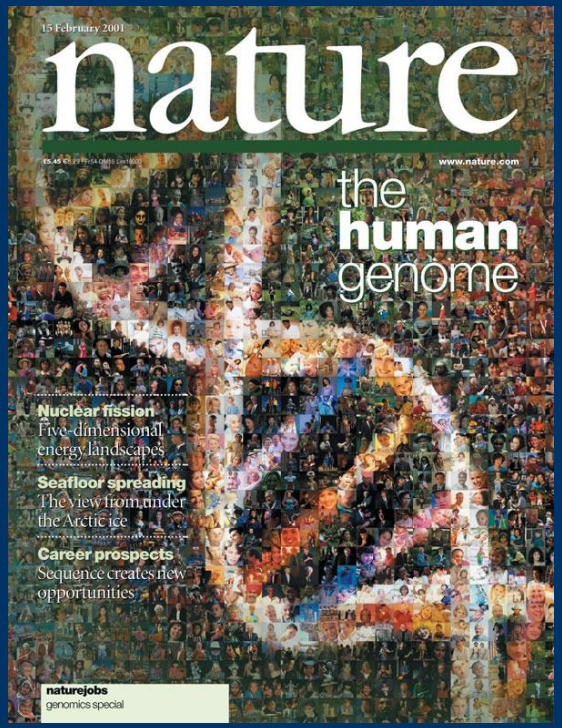
The 1000 Genomes Project Data Tutorial

November 3, 2010

2001


15 February 2001
nature
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

2010


nature
A THOUSAND GENOMES

International project to make a next generation reference data set for human genetics.

Consortium with many platforms, centers, research groups, funders.

Provide a resource to support GWAS and disease studies, in many populations.

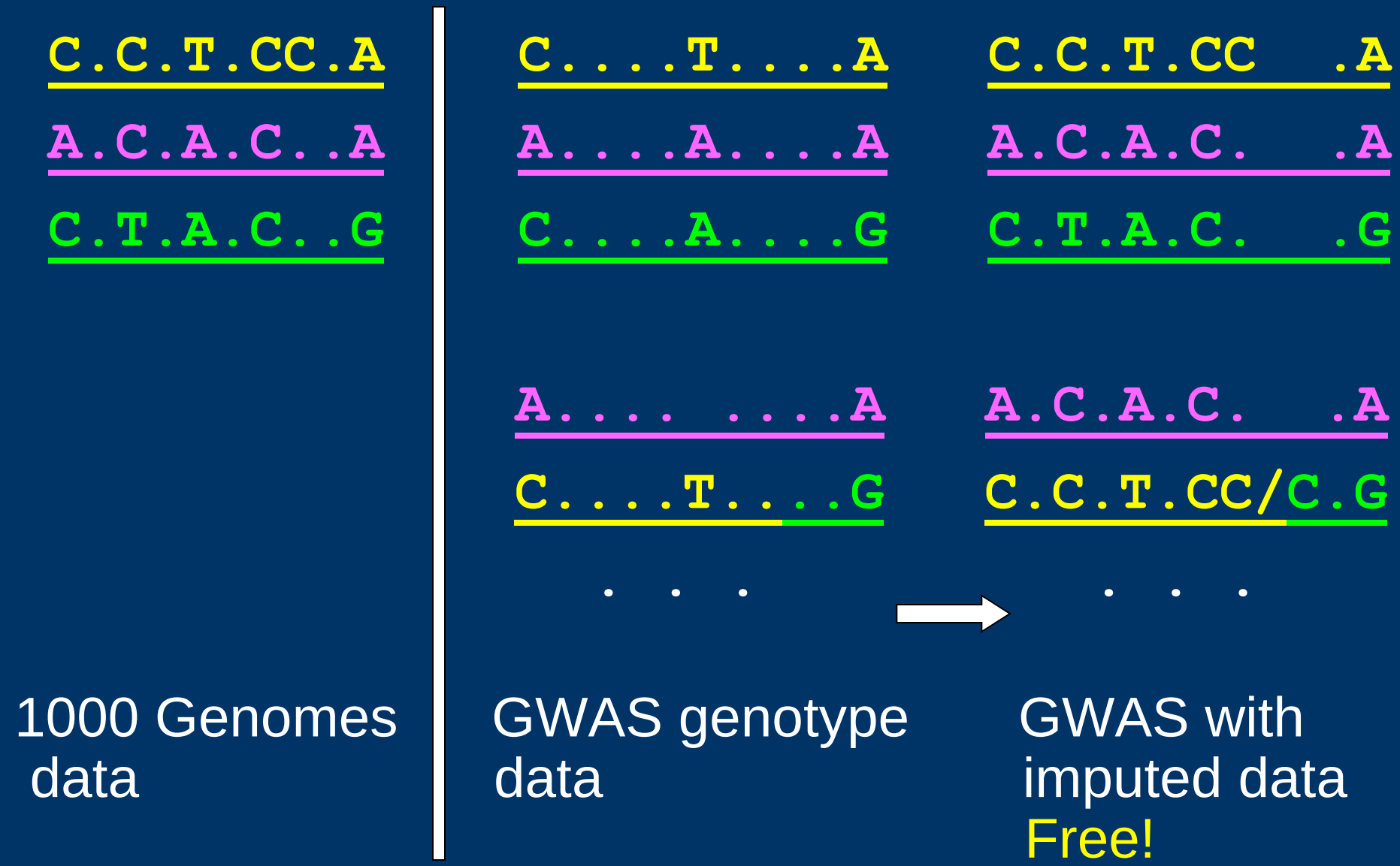
Goals

- Find > 95% of the SNPs with a frequency at least 1% (towards 0.1% in coding regions).
- Find short indels and larger structural variants.
- Provide genotypes, haplotypes, by sample.
- Provide cell lines for the samples.
- Release the data publicly and quickly.

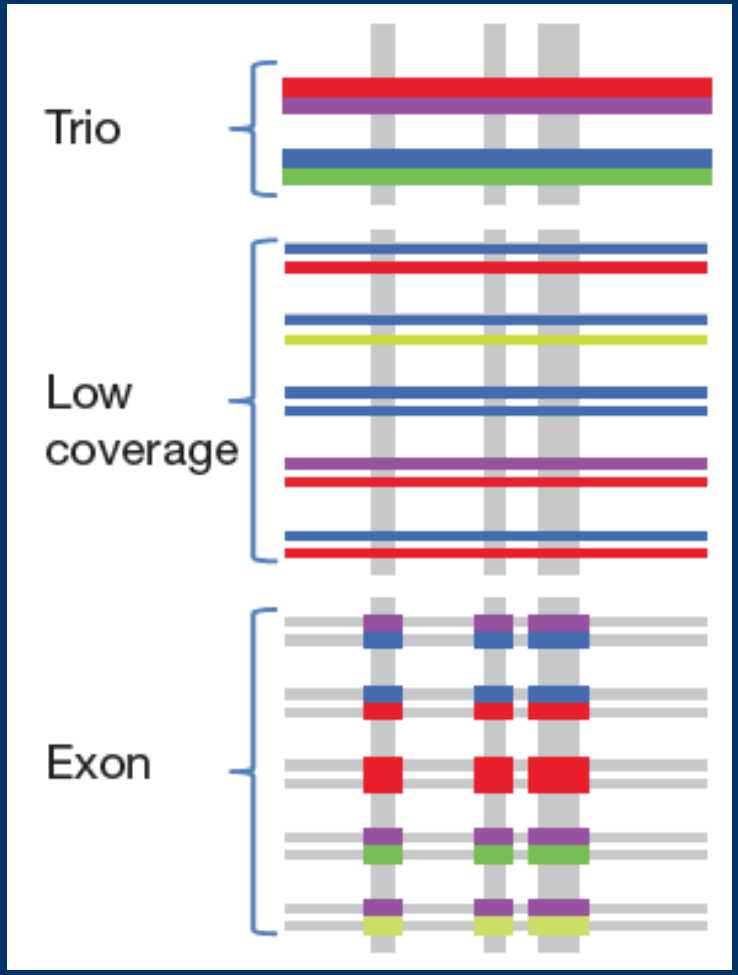
Uses of the Resource in Medical Genetics

- Impute untyped variants in GWAS, to better find association regions.
- Look up most variants in regions of GWAS hits - the set of possible causal variants.
- Prioritise and filter variants when looking for new mutations in Mendelian diseases.

Imputing Untyped Variants



Three pilot sequencing strategies



Populations	Samples	Coverage
2	6	20-40x
4	179	2-4x
7	697	20-50x

2500 Samples in Full-Scale Project

Han Chinese, Southern Han Chinese, Dai Chinese, Japanese, Kinh Vietnamese

Utah CEPH, Tuscan, British, Finnish, Spanish

Yoruba, Luhya, Gambian (two pops), Malawian

African-American in SW US, African-American in Mississippi, African-Caribbean in Barbados

Mexican-American, Puerto Rican, Colombian, Peruvian

Ahom, Kayadtha, Reddy, Maratha in India, and Punjabi in Pakistan

No identifying or phenotype data

All samples available (or will be) from Coriell

Plans for the Full-Scale Project

Sequence data

- Interim release with >25M SNPs on >600 samples today.
- 4-6X in 1100 samples (available) by Nov 2010
  - release 1st Q 2011.
- 4-6X in 600 samples (collecting) in 2010-11.
- 4-6X in 800 samples (plan collect) in 2011-12.
- Exome sequence for all of the samples.

Genotype data

- Up to 10 million SNPs with new Illumina arrays based on project data.
- Several types of structural variants in sets of the samples.

Tutorial

1. Introduction	
2. Description of the 1000 Genomes data	Gabor Marth
3. How to access the data	Stephen Sherry
4. How to use the browser	Paul Flicek
5. Structural variants	Jan Korbel
6. How to use the data in disease studies	Jeff Barrett
7. Q&A	

Video and slides **www.genome.gov** mid-Nov