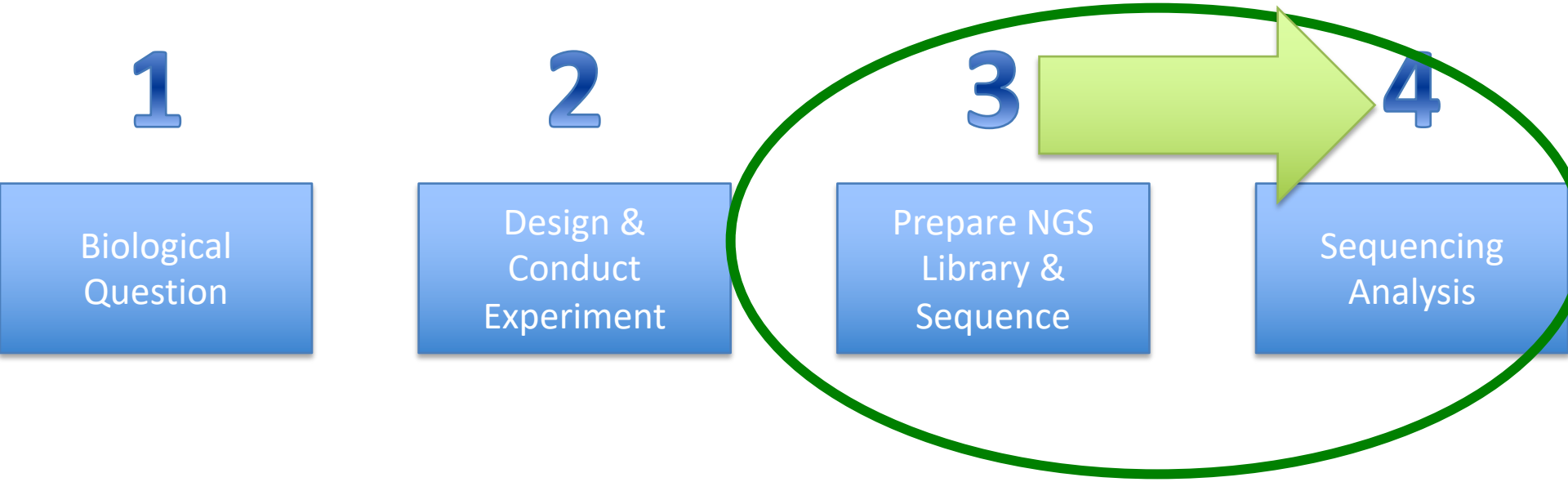
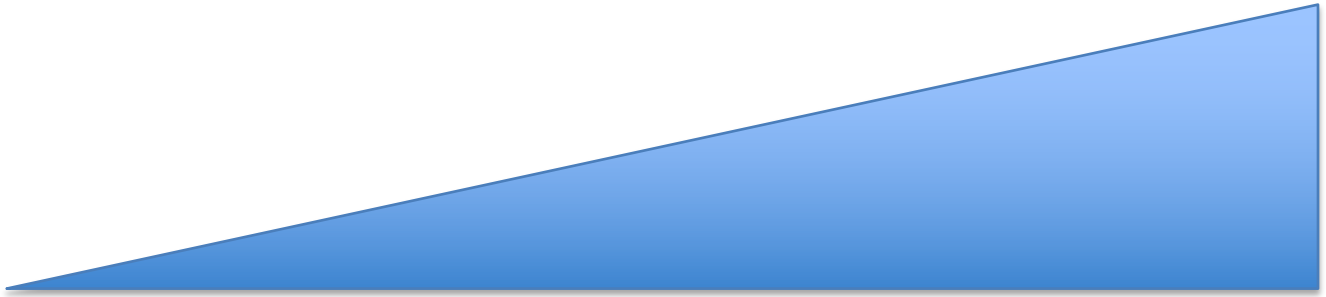
NGS Errors:

Where do they come from?
How do we get rid of them?
How do we identify them?
When do they matter?

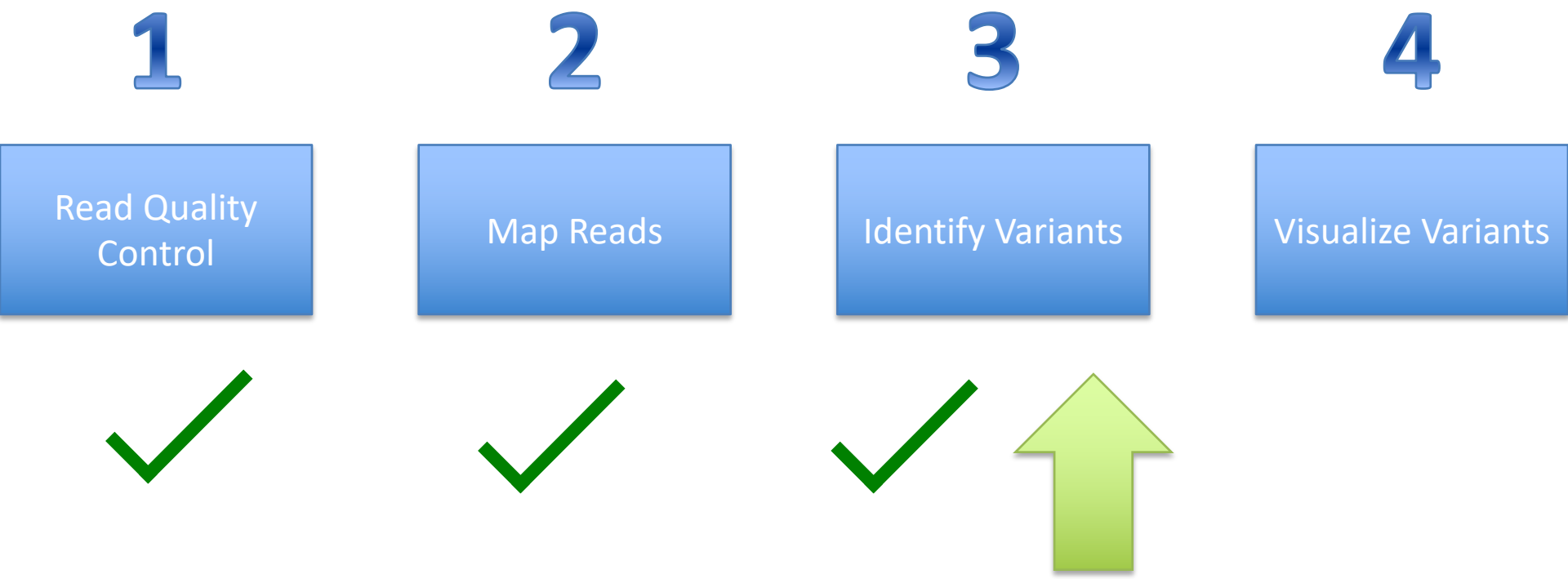
4 Main Stages



Class time



4 Typical Stages of Variant Analysis



WHERE DO ERRORS COME FROM?

Standard Illumina Library Prep

1. Harvest gDNA
2. Shear gDNA
3. Blunt-end-repair DNA
4. dA-tail DNA
5. Ligate Adapters to DNA
6. PCR Amplify Library
7. Sequence

DNA Damage source

Source of DNA	Potential Damage	Comments	References
Ancient DNA	abasic sites, deaminated cytosine, oxidized bases, fragmentation, nicks	Cytosine deamination has been reported to be the most prevalent cause of sequencing artifacts in ancient DNA.	Gilbert, M.T. et al. (2007) *Nuc. Acid Res.*, 35, 1–10. Hofreiter, M. et al. (2001) *Nuc. Acid Res.*, 29, 4793.
Environmental DNA	fragmentation, nicks (plasmid or genomic)	Nicks and fragmentation can increase the formation of artifactual chimeric genes during amplification.	Qiu, X. et al. (2001) *Appl. Envir. Microbiol.*, 67, 880.
Source of Damage			
Exposure to Ionizing Radiation	abasic sites, oxidized bases, fragmentation, nicks	Ionizing radiation is used to sterilize samples.	Sutherland, B.M. et al. (2000) *Biochemistry*, 39, 8026.
Exposure to Heat	fragmentation, nicks, abasic sites, oxidized bases, deaminated cytosine, cyclopurine lesions	Heating DNA accelerates the hydrolytic and oxidative reactions in aqueous solutions.	Bruskov, V.I. (2002) *Nuc. Acids Res.*, 30, 1354.
Phenol/Chloroform Extraction	oxidized bases	Guanine is more sensitive to oxidation than the other bases and forms 8-oxo-G. 8-oxo-G can base pair with A making this damage potentially mutagenic.	Finnegan, M.T. (1995) *Biochem. Soc. Trans.*, 23, 403S.
Exposure to Light (UV)	thymine dimers, (cyclobutane pyrimidine dimers) pyrimidine (6–4) photo products	UV trans-illumination to visualize DNA causes thymine dimer formation.	Cadet, J. et al. (2005) *Mutat. Res.*, 571, 3–17. Pfeifer, G.P. et al. (2005) *Mutat. Res.*, 571, 19–31.
Mechanical Shearing	fragmentation, nicks	Normal DNA manipulations such as pipetting or mixing can shear or nick DNA.	
Dessication	fragmentation, nicks, oxidized bases		Mandrioli, M. et al. (2006) *Entomol. Exp. App.*, 120, 239.
Storage in Aqueous Solution	abasic sites, oxidized bases, deaminated cytosine, nicks, fragmentation	Long term storage in aqueous solution causes the accumulation of DNA damage.	Lindahl, T. et al. (1972) *Biochemistry*, 11, 3610 and 3618.
Exposure to Formalin	DNA-DNA crosslinks, DNA-protein crosslinks	Formaldehyde solution that has not been properly buffered becomes acidic, increasing abasic site formation.	Workshop on recovering DNA from formalin preserved biological samples. (2006) The National Academies Press.

From NEB expressions Spring 2007, vol 2.1
By Thomas C. Evans, Jr., New England Biolabs, Inc.

Link

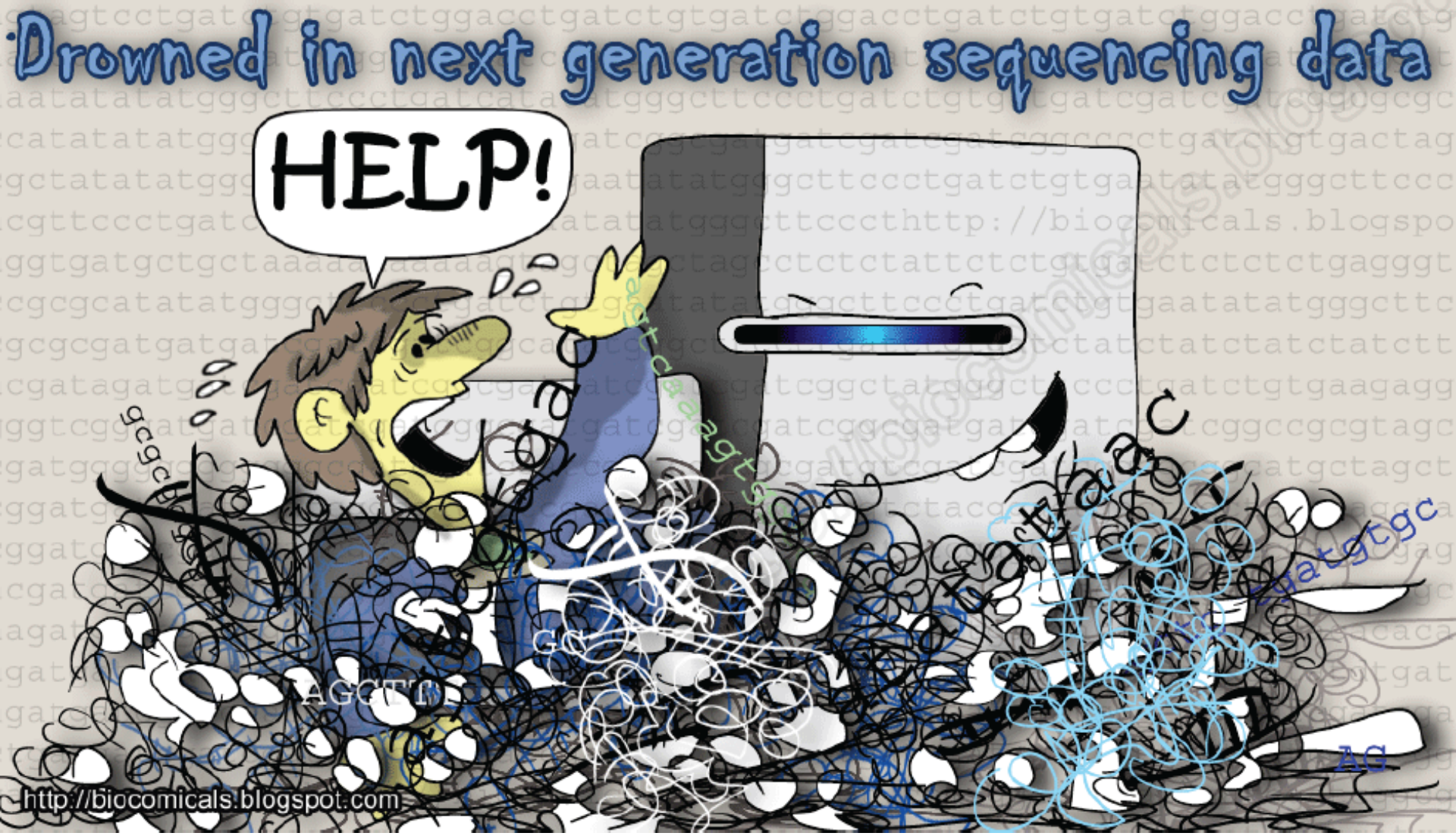
Sources of Errors in Illumina Library Prep

1. Harvest gDNA
2. Shear gDNA
3. Blunt-end-repair DNA
4. dA-tail DNA
5. Ligate Adapters to DNA
6. PCR Amplify Library
7. Sequence

1. Outgrowth, Storage
2. DNA Oxidation
    1. Costello et al 2013 NAR
3. T4 DNA Pol est. $1*10^{-5}$
    1. Keohavong, Thilly 1989 PNAS
4. Interactions with ligation?
5. ~11% 5` anti-T pro A/G bias
    1. Seguin-Orlando et al 2013 PLOS ONE
6. Phusion $4.2*10^{-7}$
    1. Li et al 2006 Nature Methods
7. Sequence specific, PCR
    1. Nakamura et al 2011 NAR

Final Results

- Overall error rate estimated between 0.1 and 1% per base (Lou et al 2013 PNAS).
- That's 1 error per 100-1000 bases sequenced, or typically at least 1 error per 3-4 paired reads.
- Up to 2 Billion 150bp PE reads / run means between 600 million and 6 billion errors per run!
- Minimum detection limit is between 0.1 and 1%



IF EVERYTHING WE DO GENERATES ERRORS WHAT CAN WE DO?

Some Suggestions to Minimize Errors

1. Minimize sample handling after biological relevance.
2. Minimal PCR whenever possible.
3. Pay attention to directionality of reads supporting variant.
4. Make use of quality score information.
5. Use sequence specific error profiles to eliminate false positives.
    1. Meacham et al 2011 BMC Bioinformatics
    2. Ma et al 2019 Genome Biology
6. Leverage other biological knowledge whenever possible (ie timecourse data).

A Note On False Negatives

- CNV between repeat elements can be virtually invisible (particularly at low levels).

Practical Limitations

- If planned sequencing coverage is less than ~100, most not important (except sequence specific). Always assume something seen once or few times is not real.
- If looking for phenotypes, (driver mutations in cancer, other disease associated mutations, antibiotic resistance, etc), causal mutations not likely to be rare.
- Massaging standard Illumina data is likely to be less effective than better experimental planning and design using alternative library preparation methods.

HOW DO WE IDENTIFY THEM?

"I HATE TO ASK AGAIN, BUT WHAT IS IT I'M SUPPOSED TO BE LOOKING FOR?"

Characteristics of True positives

Table 1. Empirically derived filtering parameters for putative somatic mutations

Parameter	Description	Requirement
Read position	Average variant position in supporting reads, relative to read length	Between 10 and 90
Strandedness	Fraction of supporting reads from the forward strand	Between 1%–99%
Variant reads	Total number of reads supporting the variant	At least four
Variant frequency	Variant allele frequency inferred from read counts	At least 5%
Distance to 3′	Average distance to effective 3′ end of variant position in supporting reads	At least 20
Homopolymer	Number of bases in a flanking homopolymer matching one allele	Less than five
Map quality difference	Difference in average mapping quality between reference and variant reads	Less than 30
Read length difference	Difference in average trimmed read length between reference and variant reads	Less than 25
MMQS difference	Difference in average mismatch quality sum between variant and reference reads	Less than 100

Koboldt et al 2012 Genome Research

Knowing Limitations of Analysis

Mutation Type	Single End	Paired End	Mate Pair
SNV	Yes	Yes	Yes
Mobile element insertion	Yes	Yes	Yes
Duplications	Unlikely	Yes	Yes
Inversions across repeat elements	No	Unlikely	Yes
SNPs in repeats	No	Limited	Yes
Insertion of novel sequence	No	No	No

WHEN DO THEY MATTER?

Assuming you believe the mutation Is real … Does it matter?

- "Common" mutations often don't matter.
  - dbSNP – humans
  - Other experiments
    - topA, spoT, pykF – Long Term Evolution Experiment
    - Organism/process specific
- More disruptive the mutation, the more likely it is to be disrupting something.
  - A synonymous mutation in the 2nd codon MUCH less likely to be relevant than a frame-shift at the same location

- After that … Science
  - More experiments
  - More resources